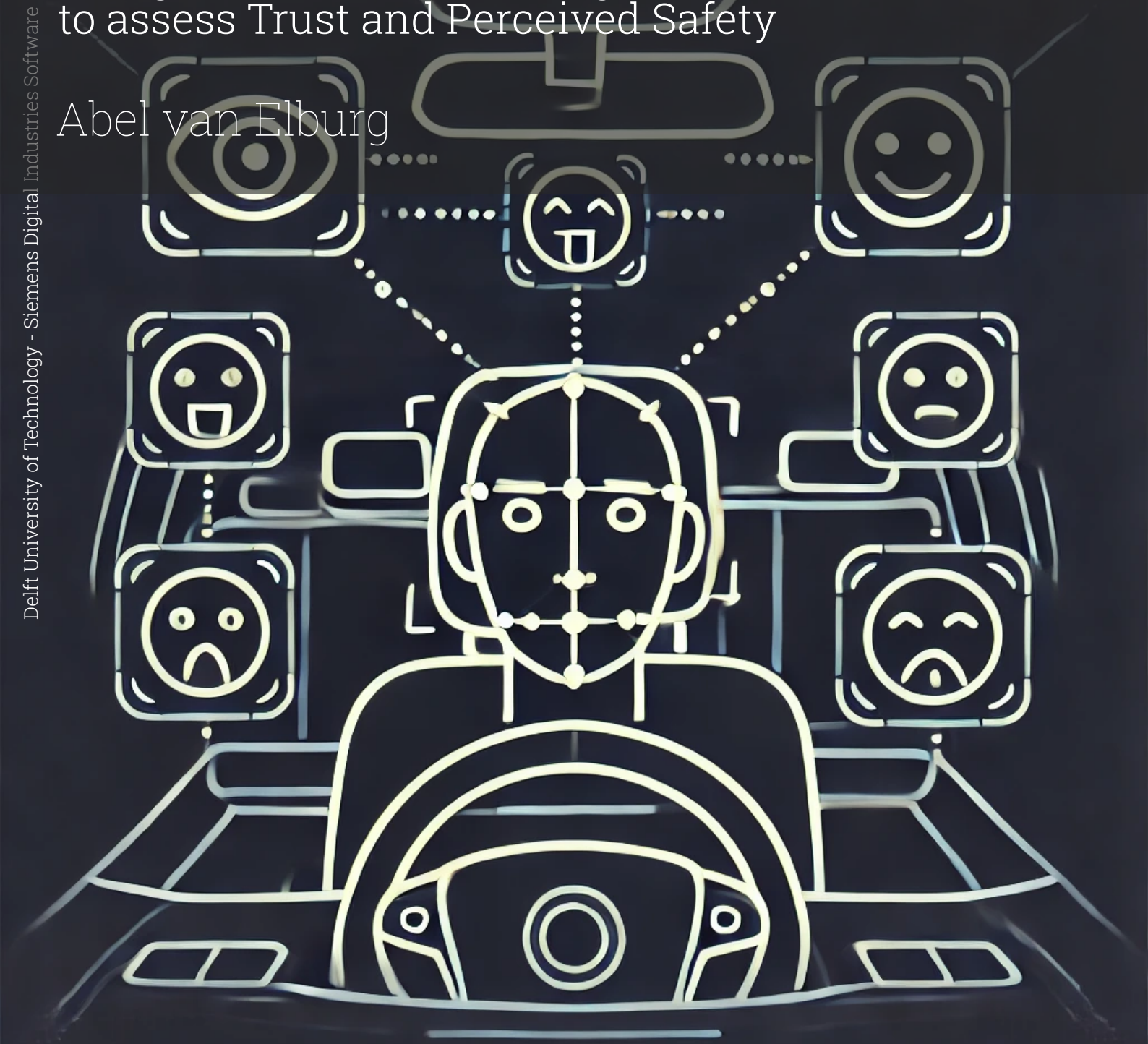
Comfort Assessment in Automated Vehicles

Using Facial Emotion Recognition to assess Trust and Perceived Safety

Abel van Elburg

Delft University of Technology - Siemens Digital Industries Software



TU Delft

SIEMENS

Comfort Assessment in Automated Vehicles

Using Facial Emotion Recognition to assess Trust and Perceived Safety

by

Abel van Elburg

to obtain the degree of Master of Science in Robotics
at the Delft University of Technology,
to be defended publicly on Wednesday August 28th 2024 at 10:00.

Student number:	4606833	
Project duration:	November 1, 2023 – August 28, 2024	
Supervisors:	Prof. dr. ir. R. Happee,	TU Delft
	Ir. V. Kotian,	TU Delft
	Ir. M. Sarrazin,	Siemens DI Software
	Ir. K. Gkentsidis	Siemens DI Software
	Dr. Ir. S. Barendswaard	Siemens DI Software
Committee:	Prof. dr. ir. R. Happee,	TU Delft
	Dr. Ir. G. Papaioannou,	TU Delft
	Ir. X. He,	TU Delft
	Dr. Ir. S. Barendswaard	Siemens DI Software

An electronic version of this thesis is available at http://repository.tudelft.nl/.

TU Delft

SIEMENS

Acknowledgements

This thesis has been an incredible journey, and it seems only right to start by expressing my gratitude to the people who have not only made it possible but also made it better by immeasurable amounts. Without the people I am about to name here, the thesis that lies before you would not have been possible, and I cannot put into words how much their guidance, support, and friendship have meant to me.

First of all this whole thesis would not have been possible without my supervisors, both from the TU Delft and Siemens. Professor Riender Happee and Varun Kotian from TU Delft, and Mathieu Sarrazin, Sarah Barendswaard, and Konstantinos Gkentsidis from Siemens. They have helped me find my way through this project at times when it seemed impossible to me. Whether it was helping me with the contents of this work itself, or just being there as someone to talk to when things did not go as planned, the support was always there and had an immediate impact both on me and on this work. I am very much aware that I might have tested their patience more than once, which makes me more grateful that they stayed with me and kept me on track.

Next, I would like to thank Flavia Acerbo from Siemens, whose knowledge of and experience with the simulator was indispensable in setting up the experiment. Without her availability and patience for all and any questions I had, whether small or big, it would not have been possible.

Then I want to thank my family, specifically my parents, not only for their support during this thesis but throughout all of my 8 years of being a student. These years have been truly amazing, and are something that I will cherish for the rest of my life. This would not have been possible without them, and I count myself extremely lucky.

I also would like to thank Peter van der Putten, assistant professor in Leiden and a university friend of my father, for proofreading this thesis. Your comments helped me dot the i's and cross the t's.

Finally, I want to mention my fellow interns at Siemens, who transformed this experience into so much more than just an academic one. Thanks for everything!

Abel van Elburg
Delft, September 2024

Summary

The advancements in automated driving hold the potential to improve our daily lives by reducing traffic accidents, minimizing congestion, and allowing more time to spend on other activities. However, realizing these benefits depends heavily on widespread public acceptance, which, according to current research, remains low. Trust and comfort have been indicated as key factors to increase acceptance, highlighting the importance of finding methods of measuring these concepts. This thesis contributes to the development of objective methods for evaluating the emotional state of drivers and passengers in automated vehicles.

In this thesis, an experiment was designed and conducted using a driving simulator to allow participants to experience a ride in an automated vehicle. The aim was to elicit varying levels of comfort by altering the driving style and introducing the presence of a pedestrian, while simultaneously collecting a comprehensive dataset to analyse these comfort levels. This dataset includes continuous subjective comfort ratings given by the participant, vehicle dynamics from the real-world drive on which the simulation was based, webcam footage monitoring the person's facial expression, Galvanic Skin Response, heart rate (variability), and eye-tracking from 32 participants. Such comprehensive datasets are rare in the literature and provide valuable opportunities for future research to compare different signals and explore their interrelations.

From the subjective comfort ratings, it was found that the driving style was a bigger factor than the presence of a pedestrian. Even though the pedestrian did cause a decrease in comfort, the difference between the two driving styles was found to be significantly bigger.

For facial expression recognition, a state-of-the-art model was successfully implemented. With minimal lighting conditions, the face could always be detected, and expressions were successfully classified with corresponding emotion labels from the universal set of emotions. Out of the 32 participants, 24 were included in the analyses. Most of these participants (15/24) did not show any detectable reaction in their facial expressions to the critical event. Amongst the 9 participants who did, 8 of them showed a Happy expression, and only 4 showed a Surprise expression. Fear was never dominant. This result shows that, in the current experiment, Facial Expression Recognition is not a reliable method for discomfort detection in Automated Vehicles.

Additionally, a neural network was implemented to predict a person's subjective comfort based on vehicle dynamics and their Galvanic Skin Response (GSR). The model was validated using Leave-One-Out Cross-Validation (LOOCV), where one participant was excluded from the training set and their data was used for testing. The results were promising, as the self-reported comfort and the model-predicted comfort showed a positive correlation across all participants. These findings demonstrate the potential for objective comfort assessment in automated vehicles, reducing the biases inherent in subjective evaluations and paving the way for further research in this field.

Contents

1

Introduction

Autonomous Vehicles, or self-driving cars, have already made appearances in numerous movies and television shows, and with the developments in automated features becoming more common, it is believed that the fully automated car will further appear on the road in the upcoming decades (Wadud et al. 2016; Winkel et al. 2023; Harb et al. 2021). In some places, it is even already possible to use driverless taxis. In San Francisco, it is possible to book a robot taxi from Waymo. In August 2024 they report to have 100.000 drives per week (Vleugel 2024). Next to enabling drivers to engage in non-driving related tasks (NDRTs), it is also predicted that it will vastly reduce traffic accidents and congestion, as the majority of these accidents are still attributed to human error (AbuAli et al. 2016).

However, apart from technological developments, these potential benefits depend on a high level of acceptance and usage by the public (Bellem et al. 2018; Bhide et al. 2023). As was very well stated by Waytz et al. 2014: "Even the greatest technology, such as vehicles that drive themselves, is of little benefit if consumers are unwilling to use it.". This is arguably the biggest obstacle that the automotive industry has to overcome, as research has shown a generally negative attitude towards automated vehicles (Su et al. 2023; Park et al. 2022; Mara et al. 2022).

According to Paddeu et al. 2020, trust and comfort are the most important factors for the public to accept this new technology. They found a strong correlation between passenger comfort and trust in an experiment with an autonomous shuttle vehicle. Another study on the factors influencing the use of partially automated driving indicated that lack of trust was the primary reason for people not to use the automated systems that are currently present and allowed in vehicles (Nordhoff et al. 2023). Trust is found to be heavily correlated with perceived risk (He et al. 2022). They also found that prediction models could be used interchangeably for these concepts. These findings confirm the intuition that when a person trusts something or someone that they might need to rely on, in this case, the automated vehicle, they will perceive less risk, and overall will feel more comfortable using it. The idea, again intuitive, is that when more people feel comfortable relying on automated vehicles, the acceptance will grow which can bring us towards the promising benefits listed above.

Comfort, and more specifically trust and perceived safety, are key factors in fostering wider public adoption of automated vehicles. Historically, research on driving comfort has focused predominantly on manual driving, with limited exploration of passenger comfort in the context of automated vehicles (Bellem et al. 2018). The transition to automated vehicles, where the driver becomes a passenger, highlights driving style as a crucial element influencing perceived comfort. It is noted that elements like passive safety systems and seating arrangements may not undergo significant changes even as automation levels increase (Bellem et al. 2018). It is therefore important that more research is done on these concepts in automated driving. Different methods are used and investigated to assess these concepts, which has proven a challenge due to their subjective properties. This thesis contributes to this area of research by collecting a comprehensive dataset and analyzing subjective comfort ratings, facial expressions, and galvanic skin responses (GSR). GSR has proven to be a reliable indicator for arousal and stress (Dawson et al. 2007; Jaiswal et al. 2023). This correlation has also been confirmed in the context of driving (Wang, Murphey, et al. 2019; Memar et al. 2021), making it a particularly valuable signal to collect both for this thesis and for future research.

It is also proposed that performing autonomous vehicles passengers' emotion detection will improve

the comfort and trust of the passenger (Sini, Marceddu, Violante, Dessì 2020). The car would then be able to adapt its driving style according to the state of the passenger. This would require real-time detection, which is more challenging. The objective of this thesis is to investigate the use of facial emotion recognition as a method for comfort assessment in automated vehicles, in post-processing. Facial emotion recognition is often based on the discrete set of universal emotions defined by Paul Ekman (Ekman et al. 1999). These emotions are expressed with the same facial expressions for all humans, and this set consists of Anger, Fear, Disgust, Happiness, Surprise, and Sadness. Especially Fear and Surprise are of interest to detect in this context, as these would indicate a person feeling unsafe or uncomfortable. This makes facial emotion recognition a relevant method to investigate in the context of comfort assessment in automated driving.

Comfort is influenced by different factors. The main factors are trust and perceived safety, which are heavily correlated, and motion sickness. In this thesis, the focus was on trust and perceived safety. The research question is as follows: How can facial emotion recognition benefit the assessment of comfort, concerning trust and perceived safety, in automated vehicles?

To answer this question, an experiment was conducted in a driving simulator. The participants were subjected to a scenario as if being driven by an automated vehicle. The scenario varied in driving style, and was with or without a pedestrian in the scene, to incite different levels of comfort. While giving a continuous subjective comfort rating via a knob, a camera recorded their face, and other physiological data was collected as well. After the data collection, the data is analyzed, investigating possibilities for predicting the given comfort rating from the available data. Apart from the research question, the following hypotheses are formulated:

- H1: Including facial emotion recognition in the dataset will improve the performance in terms of correctly predicting the subjective individual comfort level, compared to only using the vehicle data and GSR. Because comfort is subjective, it differs per person per situation, and the vehicle data does not contain this information.
  - H1.1: The biggest difference in performance will be seen in extreme cases, the cases that diverge the most from the average. With this, it is meant that for participants that are overall much more or less comfortable than the average, the performance will increase the most by including facial emotion recognition.
- H2: Events that cause most subjective discomfort, will show the biggest response in facial expressions.
- H3: Fear and surprise will have the strongest, negative, correlation with feeling comfortable. When the vehicle behaves not as expected or in an unsafe way, these are the emotions expected to be seen, and the comfort is expected to decrease.

A high-level overview of this thesis is given in the diagram in Figure 1.1. First, a literature study is conducted on theoretical background and related works, which are presented in Chapter 2. This includes studies on the concept of comfort and trust in automated vehicles, and also some background information on neural networks. In Chapter 3 the methodology of this thesis is explained, motivating the choices that are made both for the data acquisition and the data processing. This gives all the information needed for the experiment itself, the main part of this thesis, which is described in Chapter 4. In this chapter, the setup and procedure of the experiment are presented. The experiment has been completed, and in Chapter 5 the data is analyzed and results are presented. Finally in Chapter 6 a discussion is given on the results, followed by recommendations for future work.

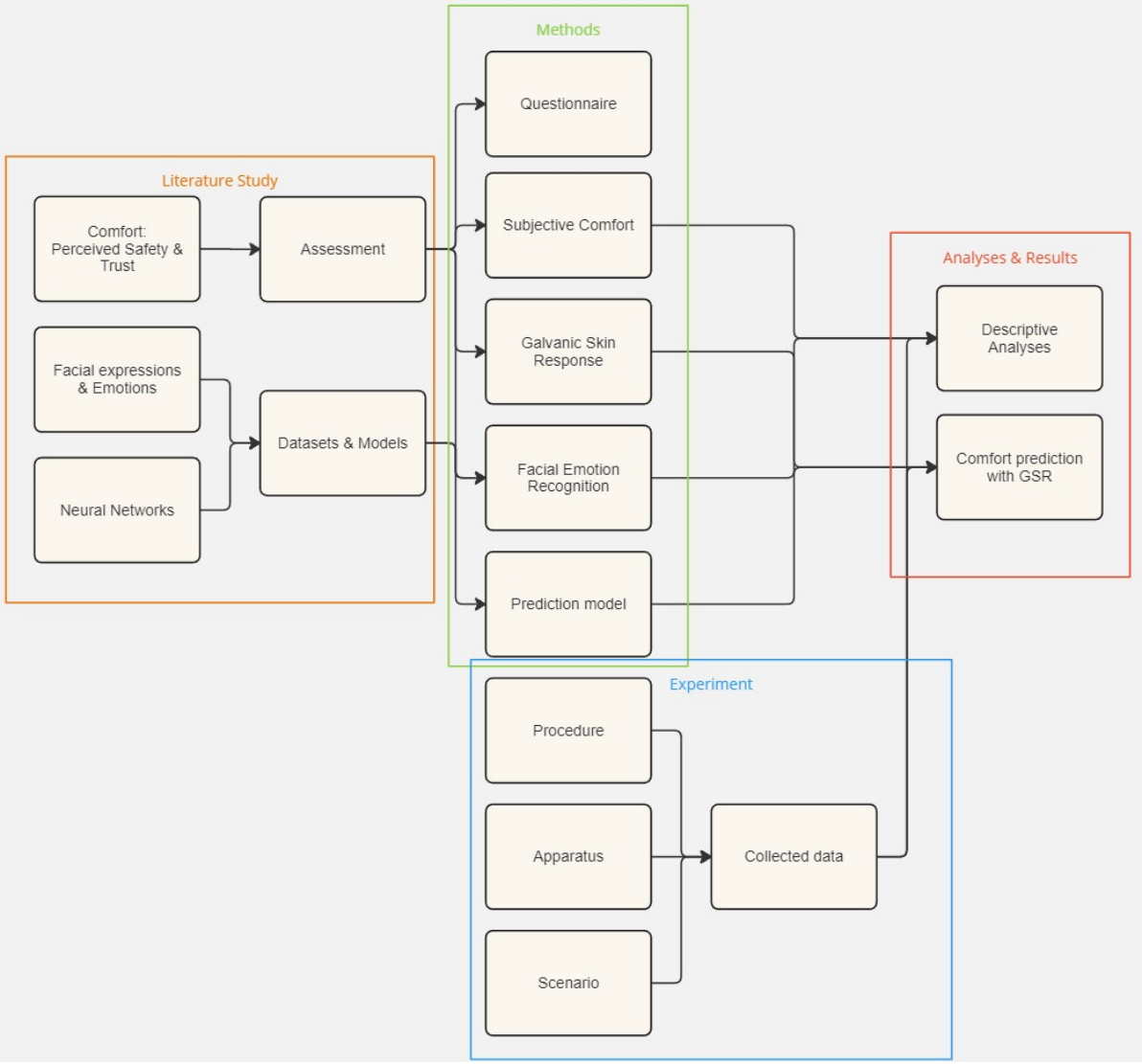


Figure 1.1: High-level diagram of this thesis. Each colored block represents a chapter in this thesis, and each sub-block is a section within that chapter. The arrows show how the different chapters are connected.

2

Theoretical background

A literature study has been conducted to acquire the prerequisite knowledge for this thesis. In this chapter, some concepts and definitions that are relevant for this thesis are presented. From the high-level overview presented in the introduction in Figure 1.1, this chapter zooms in on the part depicted in Figure 2.1. Literature on the concepts of facial expressions and emotions (section 2.1), and comfort and trust (section 2.2) was studied to avoid misunderstandings on their definitions in the context of this thesis. The current assessment methods, and their limitations, are summarised in 2.3. A concise introduction to neural networks was given in section 2.4, focusing on specific types that are used in this thesis. Finally, in section 2.5, the research for selecting the model to be used for facial expression recognition is presented.

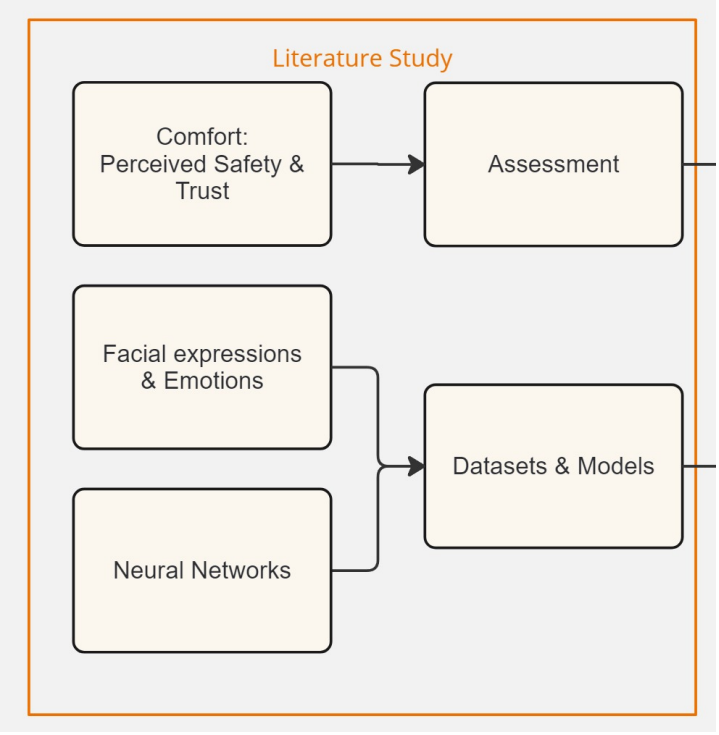


Figure 2.1: Overview of the theoretical background chapter.

2.1. Facial expressions & emotions

As humans, we can intuitively read a human face. We can see someone is sad or angry, without really studying this. Even if you have never met this person, you can likely tell how a person is feeling

by looking at their facial features. To put a label on these facial features that we use to determine a person's emotional state, Paul Ekman proposed the Facial Action Coding System (FACS) (Ekman, Friesen 1978) in 1978, which is widely used in the field of computer vision (Davoli et al. 2020). A 2020 publication of the book "What the Face Reveals", of which the first edition was released in 1997, gives an overview of what FACS is and how it has continued to stay relevant over the course of decades (Rosenberg et al. 2020). It is important to note that FACS is not a system to code emotions. It is a system to code observable actions that occur in the face, based on muscle movements. The system consists of Action Units (AUs), which represent the actions of one individual muscle, or a combination of muscles that tend to act together. The first 10 of these are shown in Figure 2.2. The attributions given to these movements, such as facial expressions or emotions, happen after the coding and are not part of FACS itself (Rosenberg et al. 2020). For example, raising the corners of the mouth is perceived as smiling, and smiling is then taken as an indication of someone's emotional state, in this case Happy.

AU	Description	Facial muscle	Example image
1	Inner Brow Raiser	*Frontalis, pars medialis*	
2	Outer Brow Raiser	*Frontalis, pars lateralis*	
4	Brow Lowerer	*Corrugator supercilii, Depressor supercilii*	
5	Upper Lid Raiser	*Levator palpebrae superioris*	
6	Cheek Raiser	*Orbicularis oculi, pars orbitalis*	
7	Lid Tightener	*Orbicularis oculi, pars palpebralis*	
9	Nose Wrinkler	*Levator labii superioris alaquae nasi*	
10	Upper Lip Raiser	*Levator labii superioris*	

Figure 2.2: The first 10 Action Units of the Facial Action Coding System with a description of the action, the names of the muscles involved, and an example picture (Cohn et al., 2007).

For labeling emotions from facial expressions, one of the most considered sets of general emotions are the six physiologically distinct emotions theorized by Paul Ekman: anger, fear, disgust, happiness, surprise, and sadness (Davoli et al. 2020; Ekman et al. 1999). Ekman argues that the facial expressions that display these emotions are universal amongst all humans, no matter when or where they grew up. This approach is widely accepted, but categorizing facial expressions into discrete basic emotions is not undisputed. The argument is that emotions often blend together, making a discrete classification system inadequate (Savran et al. 2013; Leng et al. 2007). A popular alternative is the circumplex model of affect, proposed by Russell (Russell 1980). This 2D model features arousal on one axis and valence on the other, where arousal represents the state of intensity, and valence represents the pleasantness of an emotion (Davoli et al. 2020). Along the edge of this model, different emotions can be labeled that would generally correspond to that area of the model, as is also depicted in Figure 2.3.

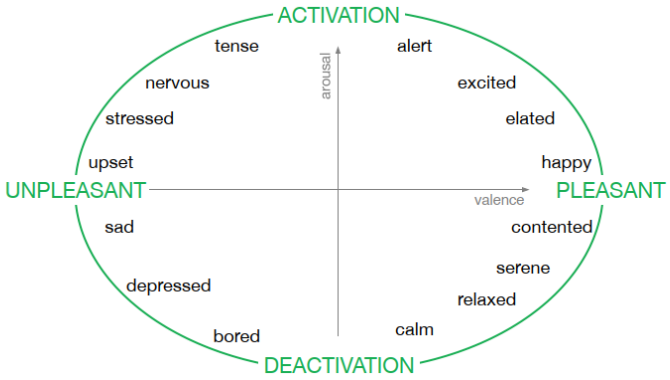


Figure 2.3: Graphical representation of Russell's circumplex model of affect (Davoli et al., 2020). On the x-axis is the valence, which represents the pleasantness of the emotion. On the y-axis is the arousal, representing the state of intensity. Along the edge, some emotions are defined that would correspond to that region in the graph.

However, when looking at the universality of expressed emotions in real-world situations, research has indicated another important factor: display rules (Ekman, Sorenson, et al. 1969; Matsumoto 1990; Safdar et al. 2009). These display rules are procedures learned to manage the expressiveness of the face for specific emotions in different situations, i.e. the displayed arousal (Ekman, Sorenson, et al. 1969). Research has shown that these rules can differ significantly between cultures, with multiple experiments showing differences between for example Japanese and North American culture when it comes to the intensity of displaying different emotions (Matsumoto 1990; Safdar et al. 2009). Interesting to note is that research on this concept also tends to make use of the basic set of emotions as defined by Ekman. So the universality of this set seems generally accepted by literature, and the core evidence for this is Ekman's research in a remote place in New Guinea (Cohn et al. 2007). People were told certain stories that would elicit a specific emotion from this set and picked between 3 pictures of faces what the corresponding expression would be (Ekman, Friesen 1971). Also, Ekman presented pictures of a man asked to display what his facial expression would be in certain situations, as can be seen in figure 2.4.

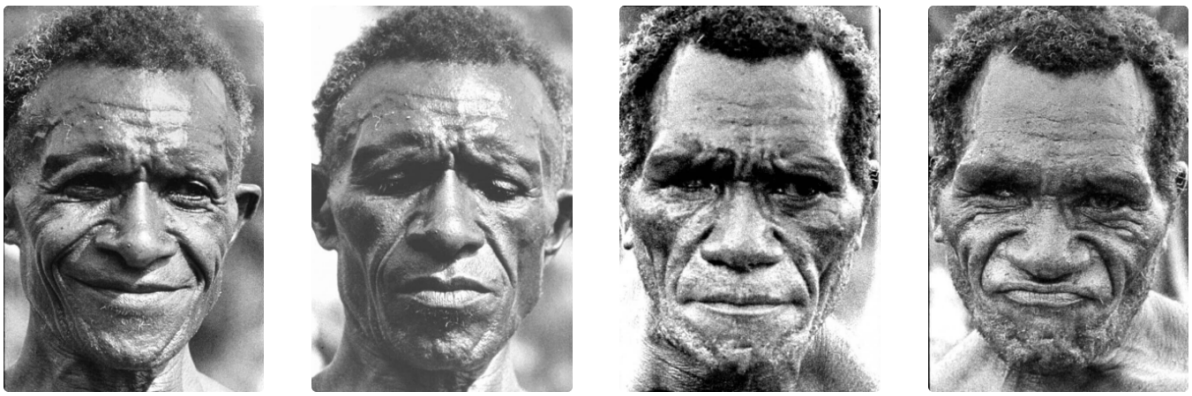


Figure 2.4: This man from New Guinea living in an isolated, preliterate culture using stone implements which had never seen any outsiders before. From left to right, Ekman asked him to show what his face would look like if: (1) Friends had come. (2) His child had just died. (3) He was about to fight. (4) He stepped on a smelly dead pig (Ekman,)

2.1.1. Emotions in driving

In the context of driving, both the discrete classification model and the arousal-valence model are used in literature. An experimental study on driver emotion recognition in 2007 (Leng et al. 2007) argues that a driver's performance is more affected by strong emotions and that the discrete classification method is more simple, direct, and convenient to evaluate. They measured different physiological signals during the experiment, while inducing certain emotions in the participants through videos, to search for correlations that could help set up a driver monitoring system that can recognize the driver's emotions. The signals were blood volume pressure, skin conductance, skin temperature, respiration rate, and

gripping force. They showed two comedic movie clips, and two scary movie clips, and then performed ANOVA analyses on the means and standard deviations for heart rate, skin conductance, and skin temperature. For the means, they conclude that both heart rate and skin conductance are significantly different, but report a P-value of 0.089 and 0.098 respectively. For the standard deviation, only the heart rate is significantly different (P=0.045). They conclude that it is necessary to develop a multimodal drivers' emotion recognition method based on physiological parameters and facial expression (Leng et al. 2007).

On the other hand, a study on modeling the emotion versus task performance relationship of drivers by Cai et al. 2011 used the arousal-valence plane instead of basic emotions. They set up virtual driving environments with different performance-related metrics such as traffic violations, braking time, and lane deviation. They find relations with performance for both arousal and valence, using self-assessment of the participants as a reference. However, for the self-assessment of the emotional state of the drivers, they did incorporate discrete emotion labels as the participants were more familiar with these. They argue that in future research physiological signals and facial expressions may help to conduct analyses on emotion-performance relationships (Cai et al. 2011).

The goal of this research is not to develop a novel facial emotion recognition method. It is to implement an existing method in the context of comfort detection in automated vehicles. In high levels of automation, the driver would essentially be a passenger. Instead of relating emotions to the performance of the driver, they would be related to the comfort of the passenger, ideally allowing the control system of the vehicle to adapt its driving when necessary. This difference makes it appealing to detect the presence of specific emotions. On top of this, during this literature review, it became clear that most published models for facial emotion recognition classify according to the discrete set of emotions, with the addition of Neutral. Therefore, for this research, it was deemed more beneficial to look at the set of discrete emotions, which the evidence suggests are universal for people from different backgrounds and cultures.

2.2. Comfort

Comfort is a concept that is widely used and chased in almost every area of research that involves human-robot interactions. However, finding one definitive definition of what describes comfort turned out to be a challenge. The ISO 5805 defines comfort as a "subjective state of well-being or absence of mechanical disturbance in relation to the induced environment" (Castellanos et al. 2014). Even though literature does not agree on one definition, one of the shared aspects of many definitions is that comfort is subjective (Bellem et al. 2018). This means that the comfort of a user is dependent on their perception. Narrowing it down to the specific context of comfort in a high-level (SAE 4+) automated vehicle, the literature points to different factors that play an important role in user comfort (Beggiato, Hartwich, et al. 2020; Peng et al. 2023), like vehicle dynamics and interaction with its environment. From the perspective of the user, perceived risk and trust in the automated system are thought to play an important role in their comfort. Motion sickness, or the absence of it, is another important concept for comfort in automated vehicles. For this research, the focus was on perceived risk and trust, and not on motion sickness.

2.2.1. Perceived safety and trust

Research has consistently indicated a generally negative public attitude toward automated vehicles (Su et al. 2023; Park et al. 2022; Mara et al. 2022). In these studies, trust emerges as a crucial factor. Even with current road-legal automated systems, a lack of trust in these systems has been identified as the main reason users choose not to use the available automated features (Nordhoff et al. 2023). Addressing and improving user trust is therefore a critical challenge for the industry, particularly if it is aimed at popularizing autonomous systems (Park et al. 2022).

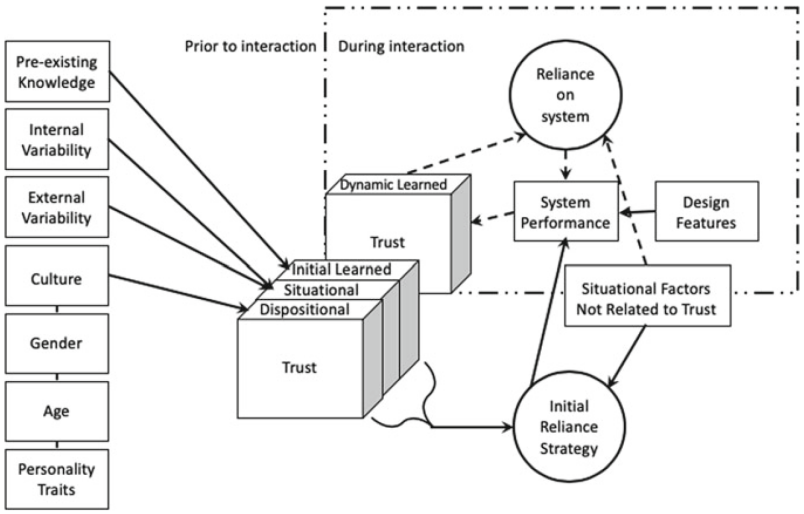


Figure 2.5: Model of trust in automation (Hoff et al., 2015).

Models of trust in automation are discussed in (Holthausen et al. 2022). Among others, they discuss a modern and comprehensive model developed by (Hoff et al. 2015). This model can be seen in Figure 2.5. In this model the authors made a distinction between the time before using a system and what happens during the time of use. They divide the trust prior to the use of the system in three categories: Initial, situational, and dispositional trust. The last one is considered a personal trait of someone that can generally not be influenced. Situational trust is about the use-case of the system and the complexity. The factors influencing can be divided in internal and external factors. Finally, the initial learned trust is based on knowledge that the user has prior to using the system. This learned trust is also the trust that directly changes during the use of a system, for example driving in an automated vehicle (Holthausen et al. 2022).

A recent study attempted to create a conceptual framework for comfort in automated driving (Peng et al. 2023). They did so by first creating a framework from literature, and then held an online workshop with experts in the field of comfort in automated driving. Part of their discussion was about what the difference was between being driven by a taxi/bus/train versus an SAE level 4 automated vehicle. A few of the highlighted differences are the duration of use, the expectations about the driving style, and the privacy. The refined conceptual framework for user comfort in automated driving can be seen in Figure 2.6. We see Trust and Perceived Safety close together. Except for the link from trust to perceived safety, both have identical relationships to the rest of the framework, specifically the physical layer and the newly added concept of Expectation. Experimental research also has shown that trust and perceived safety are directly linked (He et al. 2022), which is in line with this framework. It is noted by one of the experts -the paper does not specify which one- that expectation is prominent in the Unified Theory of Acceptance and Use of Technology Model (UTAUT) (Venkatesh et al. 2012) and the Technology Acceptance Model (TAM) (Marangunić et al. 2015), underlining its importance (Peng et al. 2023).

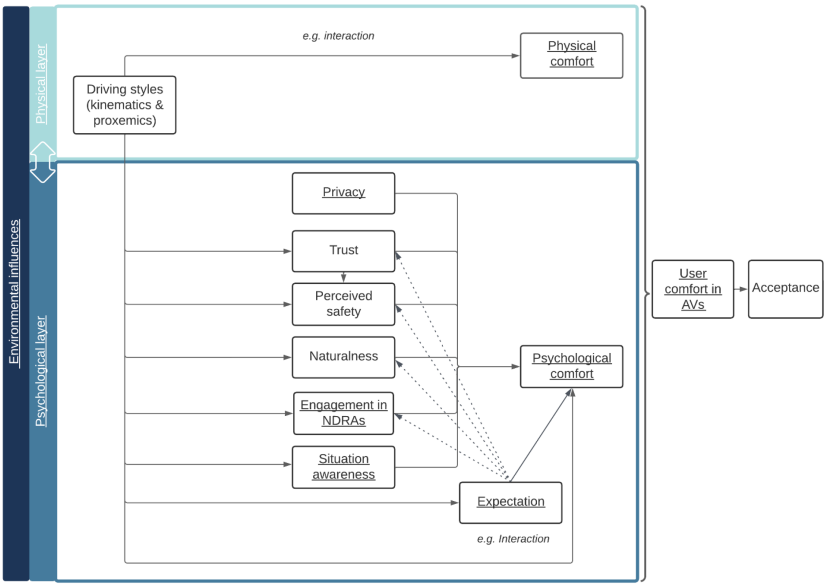


Figure 2.6: The refined conceptual framework of user comfort in automated vehicles. The concepts added by the expert workshop are underlined. Dashed lines are only used to clarify relationships where lines cross, the meaning of these lines is the same as the normal line (Peng et al., 2023).

Lord Kelvin said: ”If you cannot measure it, you cannot improve it”. This is also applicable here. To increase trust and perceived safety and achieve user comfort in automated vehicles, we need to measure it. The subjective nature of these concepts can make this challenging, underlining the scientific relevance of this research. In Section 2.3 the assessment of these concepts is investigated. Current assessment methods and their limitations are discussed, showing the gap that this work can fill.

2.3. Assessment

The measurement of trust and comfort in automated vehicles is predominantly subjective, often assessed through questionnaires (Bellem et al. 2018; Bhide et al. 2023; Kyriakidis et al. 2015; Paddeu et al. 2020; Castellanos et al. 2014; Haboucha et al. 2017; Golding 2016). However, questionnaire-based research is subject to various biases, which according to (Bhide et al. 2023) is also evidenced in studies on the acceptance of automated vehicles. They find the following biases occur in studies relating to Automated Vehicle Acceptance:

- Social desirability bias: Where respondents answer what they feel is desired, instead of what is the truth (Holtgraves 2004)*.
- Cognitive dissonance bias: Where a person’s actions do not align with their beliefs, they are known to adjust their viewpoint or behaviour (Wu et al. 2020)*.
- Hypothetical bias: When respondents give ratings without having experienced the AV (Palatinus et al. 2022)*.
- Misunderstanding: When respondents misunderstand the meaning of a question, for example when the question is formulated too generic (Bhide et al. 2023).
- Inaccuracy: When respondents do not accurately rate the question, for example when it involves their cognitive or emotional state (Palatinus et al. 2022)*.
- Response style/bias: People generally have a bias to answer questions in a certain way, for example always giving more extreme values, or the opposite, always giving more medium ratings (Jackson et al. 1958; Van Vaerenbergh et al. 2013)*.

* as cited in Bhide et al. 2023

Questionnaires also do not provide continuous measurement and often require some interruption in the experiment. To address these issues, researchers have begun using so-called direct input devices for subjective assessment (Su et al. 2023). In Beggiato, Hartwich, et al. 2020 they evaluated perceived

discomfort during automated driving with a handheld manual input device, of which a picture can be seen in Figure 2.7.

Figure 2.7: Handset control and the corresponding response scale, used to continuously assess perceived discomfort during an automated drive (Beggiato, Hartwich, et al., 2020).

The challenge that remains is that comparing subjective measures can be difficult due to some of the inherent differences listed above. This underscores the need for more objective measurements (Castellanos et al. 2014). There is an international standard ISO2631 on ride comfort, which suggests to use a frequency weighted acceleration signal, where the frequency weighting corresponds to the position and direction of the acceleration signal (Deubel et al. 2023). However, has started to focus more on person-specific signals like braking behavior and physiological signals to indicate different concepts of comfort. In (Wintersberger et al. 2016) they did both facial expression analysis, and used qualitative surveys to determine the emotional state of the participants. They compared each method with itself to see if there would be significant differences with a male, female, or Automated Driving System as driver of the vehicle.

For the concept of trust and perceived safety, research has indicated a strong and direct correlation between trust in automation and perceived risk (He et al. 2022). He et al. conducted an experiment in a driver simulator for a vehicle with SAE level 2 automation, measuring multiple physiological signals to assess their predictive value in monitoring trust and perceived risk. The signals they measured were braking signal, cardiovascular activity (ECG), and pupil dilation. Galvanic skin response (GSR) was also measured but the quality was deemed invalid and excluded from the analysis. Participants also continuously gave their subjective perceived risk ratings via a pressure sensor on the steering wheel with visual feedback. On top of that, a pre- and post-questionnaire was used to measure learned trust, and vocal feedback was asked about perceived danger and trust in the automated system during the experiment. Pupil dilation showed a correlation with perceived risk during the more risky events. Merging and braking increased the heart rate, but no quantified relation between heart rate and perceived risk was found (He et al. 2022).

The previously mentioned research by (Beggiato, Hartwich, et al. 2020) also found physiological reactions overall to work as indicators for events that cause a sudden increase in discomfort, but not so much for slowly evolving and longer lasting situations. They also found an increase of pupil diameter for uncomfortable situations, and a decrease in blink frequency (Beggiato, Hartwich, et al. 2020). For skin conductance levels they found that the location of where on the body the sensors are placed highly influences the measured effect. The same was found for the heart rate. They also did some facial expression analysis, where they found indicators for surprise and tension in close approach situations (Beggiato, Hartwich, et al. 2020).

Something else that is used to assess the emotional state of participants is electroencephalogram (EEG) (Izquierdo-Reyes et al. 2018; Park et al. 2022), which involves measuring brain signals that are related to different emotions. In Izquierdo-Reyes et al. 2018 an approach was proposed to identify the emotional state of subjects in elicited emotion experiments. A K-Nearest Neighbors algorithm using Euclidian distance was used to predict the emotional state from the power spectral density of the frequency cerebral bands. The exact working of this is outside the scope of this thesis and therefore not elaborated. Park et al. 2022 also used EEG signals to monitor brain activities relating to emotional states during a self-driving car simulator. In this study they use Virtual Reality simulations and have the participants experience different scenarios to measure initial trust, trust escalation, trust reduction, trust mutation, and trust rebuilding. This trust is measured on a 5-point Likert scale after each segment. From the EEG signals the beta/alpha ratio was analyzed, where high ratio indicated stress or anxiety, and low ratio indicated calmness or positive emotional states. A direct correlation was found between the participants trust and their emotional response. Trust levels decreased in response to stressful driving incidents, but could be partially rebuilt with calm driving behavior. The findings suggest that

making automated vehicles responsive to passengers' emotional states can improve trust and their social acceptability (Park et al. 2022). This highlights the potential benefits of assessing the emotional state of the passenger, underlining the importance of further research in this area.

The assessment of trust is one of the primary obstacles for continuing research in trust in automated vehicles (Holthausen et al. 2022). To find objective measurements for trust and comfort, more research is still required. Assessing the emotional state of the persons shows promising results, as shown by Park et al. 2022. With the development of neural networks for image processing, facial emotion recognition, or FER, could play a significant role towards passenger state monitoring.

2.4. Neural Networks

With complex non-linear signals, like physiological data or image classification, it can be difficult to use traditional modeling methods. For a large part, this is due to the feature extraction, which needs to be done by hand or programmed by an engineer (O'Mahony et al. 2020). The introduction of deep learning algorithms improved this as they are trained on the data itself, without the need to first extract features. In 1993 it was already predicted that neural networks would be very promising in the field of image processing (Pal et al. 1993), and this prediction turned out to be right (Egmont-Petersen et al. 2002). Recent years have witnessed significant advancements in deep learning architectures, including models designed for driver state monitoring systems, consistently improving and outperforming traditional methods (Zepf et al. 2020; Davoli et al. 2020; Mou, Zhou, et al. 2021). Also in facial emotion recognition, Deep learning-based methods have shown superior results compared to conventional methods, with ongoing enhancements further improving their performance (Ko 2018; Ngwe et al. 2023).

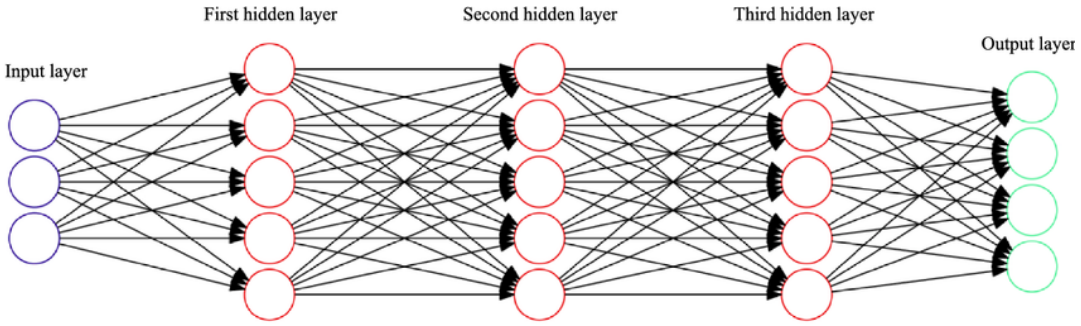


Figure 2.8: Layout of a multi-layer neural network architecture, with fully connected layers (Nedjah et al. 2019). The circles represent the nodes in that layer, and the lines represent the connections, where every connection has it's own weight. These weights are adjusted by the network during training, to get from a specific input to the desired output (Aggarwal et al., 2018).

The architecture of a neural network consists of an input layer and an output layer, with one or more hidden layers in between. The basic version is a fully connected feed-forward network, see Figure 2.8, where all nodes are connected to all nodes in the next layer (Aggarwal et al. 2018; Prince 2023). However, not all data is the same, and sometimes fully connected networks are not suitable. For this reason, different specialized types of neural networks have been developed to process specific types of data. Convolutional neural networks are optimised for grid-structured data like images, and significantly reduce the amount of parameters in a network (Aggarwal et al. 2018; Prince 2023; Goodfellow, Bengio, et al. 2016). Recurrent neural networks, more specifically Long-Short-Term-Memory networks, are designed for sequential data like time-series, extracting dependencies in time domain Yu et al. 2019. Attention layers are very popular in any type of network, allowing the network to learn which parts of the data to focus on (Vaswani et al. 2017). For more information neural networks the reader is referred to the given citations, as this thesis is not aimed at developing a new type of neural network, but rather apply existing models.

2.5. Datasets & Models

In this section public datasets are investigated and a couple of the recent works on driver state monitoring and facial emotion recognition are discussed, to illustrate how this is approached by current researchers with the development of neural networks. This thesis is not on the subject of developing

an emotion recognition model itself, the reader is referred to the original papers for more elaborate descriptions of the models. The workings of the chosen model are elaborated further in Chapter 3.

2.5.1. Datasets

Apart from the architecture of the model, the data also has a crucial role in the development of neural networks. Performance depends on good training data, and since acquiring and labeling large amounts of data can be difficult, a lot of research depends on large public datasets. In Table 2.1 an overview of some widely used datasets for facial emotion recognition can be found (Naga et al. 2023); (Li, Cui, et al. 2021). These publicly available datasets form a basis for researchers to develop and test new models. The wide use of these datasets makes it easier to compare different models by comparing their performance on the same dataset. New models can show their performance on a specific dataset and directly compare this with other state-of-the-art models that used these.

Table 2.1: An overview of some of the widely used datasets.

Dataset	Size	Participants	Environment	Labels
KDEF (Calvo et al. 2008)	4,900 images	70 participants 35 female, 35 male Age: 20-30	Static life scenario Controlled Posed	Discrete emotions: Neutral + basic set
CK+ (Lucey et al. 2010)	593 video sequences	123 participants Age: 18-50	Static life scenario Controlled Posed & Spontaneous	Discrete emotions: Neutral + basic set including contempt
RAF-DB (Li, Deng 2019)	29672 images crowdsourced and manually annotated	-	Wild	Discrete emotions: Neutral + basic set including contempt Or set of 12 compound emotions
FER2013 FER+ (Goodfellow, Erhan, et al. 2013)	35887 images	-	Wild	Discrete emotions: Neutral + basic set including contempt
EmotioNet (Fabian Benitez-Quiroz et al. 2016)	1000000 images automatically annotated	-	Wild	Discrete emotions: 23 categories of emotions
AffectNet (Mollahosseini et al. 2017)	450000 images manually annotated	-	Wild	Discrete emotions: Neutral + basic set including contempt Dimensional: Valence and arousal
DEFE (Li, Cui, et al. 2021)	164 videos of 30s	60 participants 13 female, 47 male Age 19-56	Dynamic driving scenario Controlled Spontaneosu	Discrete emotions: Neutral, Happy, Anger + 5 levels of intensity Dimensional: 9 levels of valence, arousal, and dominance

A dataset that seems worth elaborating on is the Driver Emotion Facial Expression (DEFE) dataset, tailored for intelligent vehicles. This dataset captures facial expressions triggered by video-audio clips in driving scenarios for three emotion categories: neutral, happy, and angry. For discrete classification, it is labeled with these three emotions, and also 5 levels of emotional intensity. For a dimensional model, it is labeled in 3 dimensions: valance, arousal, and dominance (Li, Cui, et al. 2021). This makes the dataset unique for emotion recognition models in automotive applications. However, it is specifically aimed at active driving scenarios. This makes it less relevant for this application where the driver is a passenger, and it is also not labeled with Fear and Surprise. The dataset is said to be publicly available, however, during this literature review, the dataset or how to access it was not found.

Finally, a dataset merging tool was developed by Jain et al. 2023. They developed a software tool called Facial Expressions Databases Classifier, which is publicly available under the MIT open-source license on GitHub (Marceddu 2020). This tool can be used to generate large facial expression picture databases by merging and processing images from various other databases. This is useful as it has been shown that even when a model can reach state-of-the-art performance on one dataset, its cross-dataset performance can be significantly less (Ngwe et al. 2023).

2.5.2. Facial Emotion Recognition model

Some of the earlier works that are specifically focused on driver emotion recognition date from 2020 (Du et al. 2020; Sini, Marceddu, Violante 2020). In the work by Sini, Marceddu, Violante 2020, they argue that the acceptance of automated vehicles will be influenced by improving the trust in these vehicles, and that to achieve this the state of the human should be objectively assessed, instead of post-experiment questionnaires (Sini, Marceddu, Violante 2020). This objective metric about comfort feeling of the passenger can be acquired through emotion analysis. They did a proof-of-concept implementa-

tion of a facial emotion recognition system using existing algorithms. Their results were promising, but the performance varied for the different emotions. The scores for contempt and disgust were lower, which was not considered a problem as these emotions are not that important in the given application, but also for fear and sadness the performance was lacking. This was partially attributed to the dataset which was known to be unbalanced (Sini, Marceddu, Violante 2020).

Deep-learning networks have been further developed for facial emotion recognition. Apart from changing the architecture, some also improve existing architectures by introducing a new loss function (Pham et al. 2023). The loss function is the value that the network tries to minimize, or sometimes maximize, during training. They present arguments on where traditional loss functions that are generally used in CNN-based FER methods have their limitations, and propose a new loss function that can overcome these limitations. Their proposed loss function is well visualized in Figure 2.9.

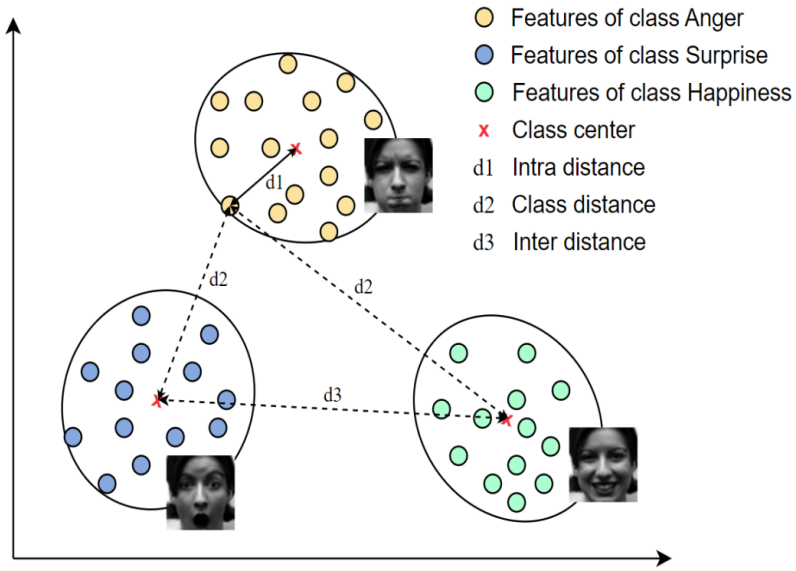


Figure 2.9: Visualization of the proposed loss function. Example given for three classes in a batch: anger, happiness, and surprise. The proposed loss function aims to reduce d1, while increasing d2 and d3 (Pham et al., 2023).

New publications on deep-learning architectures for facial emotion recognition are continuously published (Wang, Jia, et al. 2023; Chen et al. 2023; Xiao et al. 2022; Ngwe et al. 2023; Zhang et al. 2023). Wang, Jia, et al. 2023 is specifically focused on driver facial emotion recognition. They used the above-mentioned DEFE dataset to show that their proposed CNN-based architecture can outperform the previous models. An interesting finding is that they show that most wrong classifications are made between neutral and angry, suggesting that these facial expressions are more difficult to differentiate. For the model selection of this thesis both the reported performance and the availability of the code and pre-trained weights were taken into account. First, some open-source models were investigated that can be directly installed as a Python package. These models, DeepFace (Serengil et al. 2021) and FER, are both documented on GitHub. Both models perform both face detection and subsequent emotion recognition. They are based on CNNs, which are explained in Chapter 2. These models unfortunately do not have a known performance on some of the large public datasets when it comes to FER.

Another model, again based on CNNs, that was investigated was the lightweight patch and attention network (PAtt-lite) by Ngwe et al. 2023. They report state-of-the-art performance on multiple datasets and have published their code on GitHub, which included a file with pre-trained weights for the model. This pre-trained model was tested during this thesis, but when testing the model it seemed to always predict label 4 (which corresponded to Neutral). Even when using self-recorded webcam footage while performing multiple very expressive facial expressions, it would all be labeled as Neutral. When investigating this, it turned out other users also reported this problem. The developers explained that the pre-trained model was only applicable to the specific dataset (in this case the RAF-DB) that it was trained on, and that cross-database results were not good. Another user then pointed out that the de-

velopers had made an error in their code, which invalidated the claimed performance. The developers acknowledged this problem, and at the time of writing are in the process of correcting the mistake. This shows that FER is is still an ongoing area of research where new developments are made regularly. Due to platforms like GitHub developers can check each other's code and help advancements.

Another popular platform is papers with code, where the performance of published models on different datasets is tracked and compared. Different state-of-the-art models were found and investigated via this and GitHub. They were compared on performance, but also on available code, documentation, and pre-trained models. Finally, the model selected for this thesis was a Dual-Direction Attention Mixed Feature Network by Zhang et al. 2023. This model achieves start-of-the-art results on multiple datasets, and the developers released both code and the trained weights on GitHub, with the confusion matrices showing the performance on different datasets. They recently released an updated and improved version of their model under the name DDAMFN++ in the same repository, this is the model that was used. The performance on the Real-world Affective Faces DataBase (RAF-DB) (Li, Deng, Du 2017; Li, Deng 2019) can be seen in Figure 2.10. This is a dataset with almost 30-thousand images from the internet, that are annotated with emotion labels by 40 independent persons, making it an extremely valuable dataset to train and validate emotion recognition models. Overall, the performance is good, but the accuracy for Fear and Disgust are relatively low. Disgust is not expected to be relevant for this application, but Fear is. According to the confusion matrix, Fear is relatively often classified as Sad or Surprise. This can be taken into account for the data analyses. They also tested for cross-database performance, and achieved 75.6% accuracy on the RAF-DB with a model that was trained on AffectNet-7. This is a very strong achievement for cross-database performance (Zhang et al. 2023), as other state-of-the-art models like the previously mentioned PAtt-lite model struggle with this. The cross-database performance suggests this is a suitable model for our application, as it has to classify unseen data.

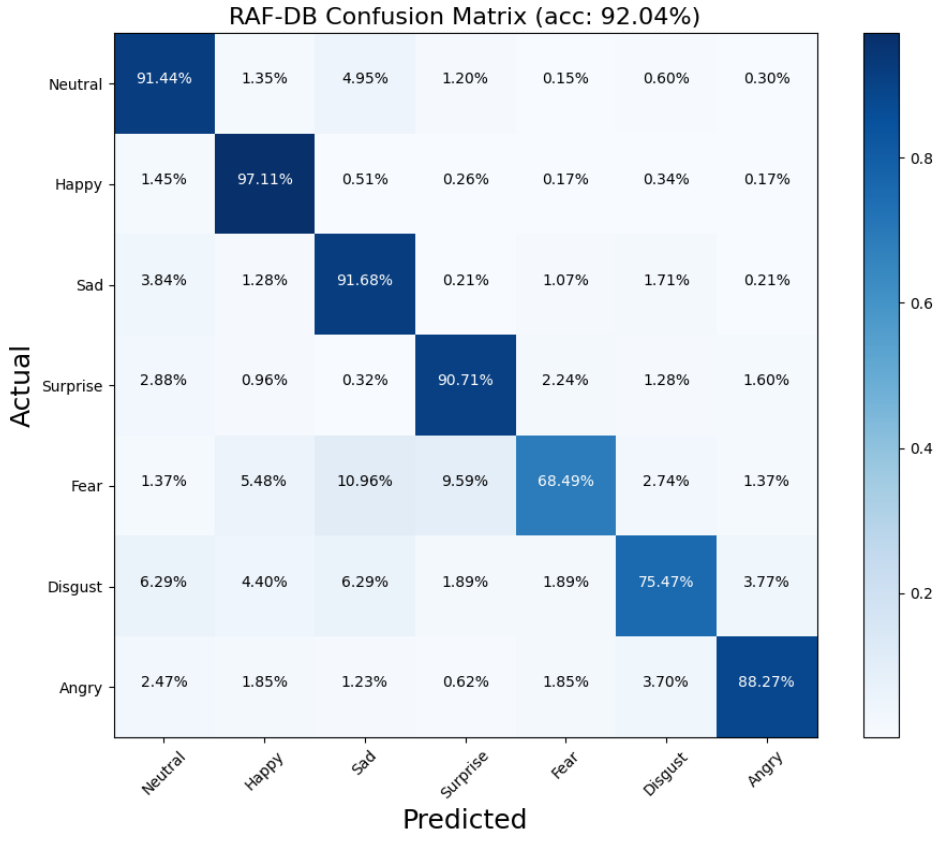


Figure 2.10: Confusion matrix for the performance of the model on the RAF-DB at the weight checkpoint that was publicly available on GitHub (Zhang et al., 2023).

2.5.3. Driver state monitoring

Apart from research in facial data or other physiological data, it is proposed that multimodal data frameworks should be used for driver state monitoring to take the context of driving into account (Mou, Zhou, et al. 2021; Mou, Zhao, et al. 2023). The work in Mou, Zhao, et al. 2023 can be considered a continuation of the work in Mou, Zhou, et al. 2021. They propose a CNN-LSTM attention-based model to fuse different non-invasive data to monitor the driver. In Mou, Zhou, et al. 2021 they classify between three stress-levels (low, medium & high). In Mou, Zhao, et al. 2023 they expand to classify emotional states in terms of valence, arousal, and dominance. As input they fuse data from non-invasive eye-tracking, vehicle dynamics, and environmental data. Their results show that combining the different data sources outperforms using them individually. Their model architecture was taken as a basis for comfort prediction in this thesis.

3

Methods

In this chapter the different methods of comfort assessment that are used for this thesis are explained. This chapter forms the bridge between the literature background from Chapter 2, and the experiment and data analyses in Chapter 4 and Chapter 5 respectively. From the high-level overview presented in the introduction in Figure 1.1, this chapter zooms in on the part depicted in Figure 3.1. The methods that will be discussed are concern processing methods for subjective comfort and GSR, an explanation of the chosen model for Facial Emotion Recognition, and the architecture of the neural network used for comfort prediction.

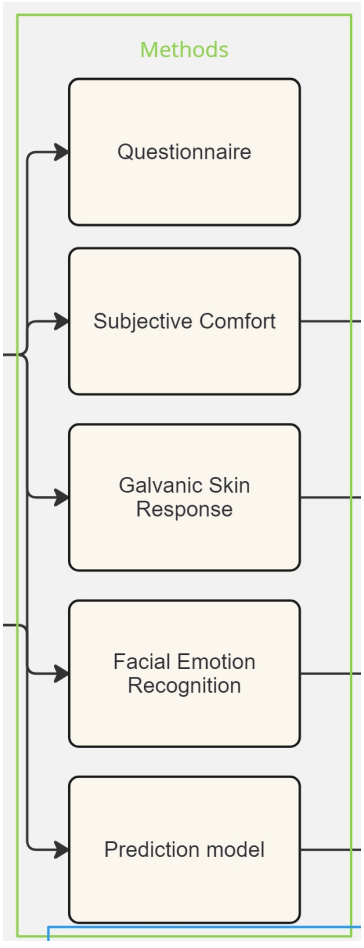


Figure 3.1: Overview of the Methods chapter.

3.1. Questionnaire

The most common method of assessing trust and or comfort is through one or more questionnaires. This method has its limitations, as discussed in Chapter 2, but it still has value, especially for assessing the participants' trust before and after the experiment. In Section 2.2, we discussed models and frameworks of trust in automation. In the model from Hoff et al. 2015 (Figure 2.5), trust was split in the trust prior to the interaction, and dynamic trust during the interaction. A questionnaire was set up to get information on the participants' pre-existing trust. The questions in this questionnaire were chosen from the work by Weigl et al. 2021, who did a literature study and a focus group with experts in the field, and then iteratively developed the questionnaire on acceptance of automated driving. They did so for SAE level 3 and level 5 separately. For this research, 8 questions were selected from their questionnaire for SAE level 5.

The questionnaire, which was conducted via Google Form, can be found in Appendix B. Each question was given as a statement, to which the participant would give their answer on a 5-point scale from 1 (fully disagree) to 5 (fully agree). After each question, the option was given to explain their answer. Similar to what other researchers have done (He et al. 2022), the same questions were presented after the experiment to see if something had changed. They were asked to elaborate only in case something had changed. The results of these questionnaires can be analyzed to assess whether an experience like this can change people's perspective on automated vehicles, and if so, how.

Before the experiment also some demographic data was collected. The participants were asked for their gender and age, if they had ever experienced a driving simulator before, and how familiar they were with automated driving systems.

3.2. Subjective comfort

To measure the dynamic trust during the experiment, researchers now commonly use continuous input devices (Su et al. 2023; Beggiato, Hartwich, et al. 2020). The benefit of these devices, as opposed to, for example, intermediate questionnaires, is that the experiment does not need to be paused. Other methods that are used are to ask the participant to vocally give a rating at certain moments, either indicated by the researcher or, for example, a beep, which also does not require a pause in the experiment. However, this only gives a discrete overview of the subjective change in comfort over time, as making these beeps too frequent can become very distracting. With the continuous input device, the participant can continuously give their subjective rating throughout the experiment. This rating can then be used to assess different driving scenarios, or compare different participants, but it can also be used as a baseline to validate other signals as assessment methods. Before doing so, the subjective ratings should be analyzed first.

The signals from the comfort knob will be inspected visually, looking for patterns that suggest the subjective ratings are to be excluded. One way to assess the reliability of the subjective scores is to check continuity over multiple runs. It is common practice to see if participants give a similar assessment during repetitions of the exact same scenario. This can be done visually, but a way to quantify this continuity is using a method called Dynamic Time Warping (DTW) (Keogh et al. 2005). DTW is a method used to measure similarity between time-series data, that may vary in time, speed, or both. It is generally used for applications like speech recognition and bio-informatics. Given two sequences $X = (x_1, x_2, \ldots, x_n)$ and $Y = (y_1, y_2, \ldots, y_m)$, the DTW algorithm works as follows (Salvador et al. 2007):

1. **Distance Matrix**: Compute the distance matrix D, where each element $D(i, j)$ represents the squared difference between x_i and y_j:

$$D(i, j) = (x_i - y_j)^2$$

2. **Cumulative Distance Matrix**: Construct the cumulative distance matrix C. The element $C(i, j)$ represents the minimum cumulative distance to reach $D(i, j)$:

$$C(i, j) = D(i, j) + \min\{C(i-1, j), C(i, j-1), C(i-1, j-1)\}$$

with boundary conditions:

$$C(0, 0) = D(0, 0)$$

$$C(i,0) = \sum_{k=1}^{i} D(k,0) \quad \text{for } i \in [1,n]$$

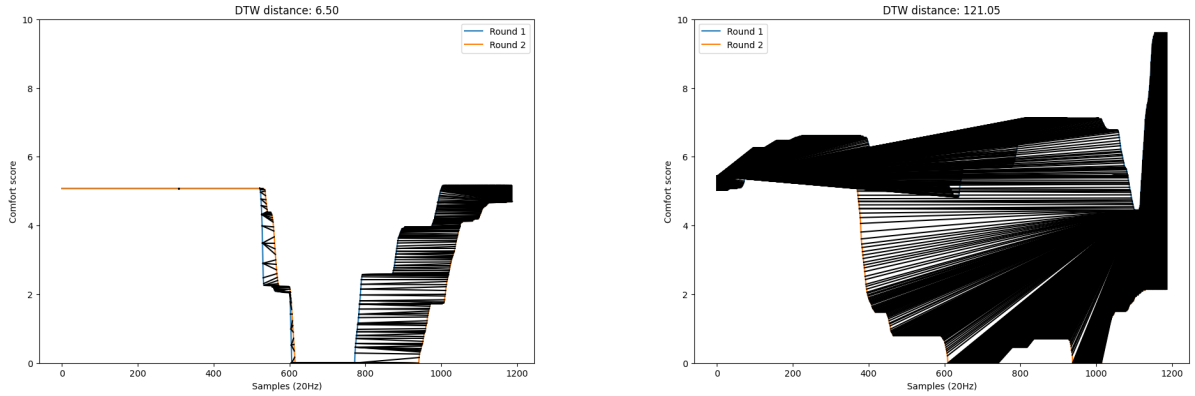
$$C(0,j) = \sum_{k=1}^{j} D(0,k) \quad \text{for } j \in [1,m]$$

3. **Warping Path**: The optimal warping path P is found by tracing back from $C(n,m)$ to $C(0,0)$:

$$P = (p_1, p_2, \ldots, p_L)$$

where $p_k = (i_k, j_k)$, $i_L = n$, $j_L = m$, $i_1 = 1$, and $j_1 = 1$.

The process is visualized for two of the runs in Figure 3.2. The left plot shows a run with high continuity, resulting in a low DTW distance. The right plot shows a run with low continuity, resulting in a high DTW distance. The paths between the two plots are drawn. The shorter this accumulated path, the more similar the two plots are. This analyses is applied to all runs and the results are presented in Chapter 5.



(a) Low DTW distance, suggesting high continuity.

(b) High DTW distance, suggesting low continuity.

Figure 3.2: Two plots showing the calculated paths between the two plots. The more distance these paths have to cover, the less aligned the two plots are.

3.3. Galvanic Skin Response

Galvanic Skin Response (GSR), Electrodermal Activity (EDA), or Skin Conductance (SC), is the signal that reflects the electrical properties of the skin. These fluctuations are caused by the triggering of sweat glands, which are controlled by the sympathetic part of the autonomic nervous system and cannot be controlled consciously. Opposed to other signals like pupil dilation or heart rate, it is not controlled by the parasympathetic system. Because of this, it is a very promising metric when looking at situations that would cause Fear. From the very beginning of research concerning EDA, it has been closely linked with emotion, arousal, and attention (Dawson et al. 2007).

For processing of the GSR data, the NeuroKit2 python package was used (Makowski et al. 2021). This is an open source package for processing physiological signals. It is completely open-source, and invites anyone to contribute. The current contributors are from all over the world and can be found with their affiliation here. They incorporate different processing methods, both self-developed and implemented from literature. They also incorporate methods from BioSPPy, a different python package for physiological signal processing (Carreiras et al. 2022). This package is marked as archived on github, whereas the Neurokit package is still being updated at this moment.

First a low-pass Butterworth filter was applied to clean the signal. As movement artifacts are likely to occur due to the dynamic nature of the experiment, the decision was made to go for a more aggressive approach, also smoothing the signal after the Butterworth filter conform Smith et al. 1997. The difference between these two approaches can be seen in Figure 3.3. The Neurokit plot is only a Butterworth filter, and the BioSPPy plot also has smoothing. Note that a small offset was given to the plots, to make

them better visible. This offset was implemented manually purely for this image, and is not present in the data itself.

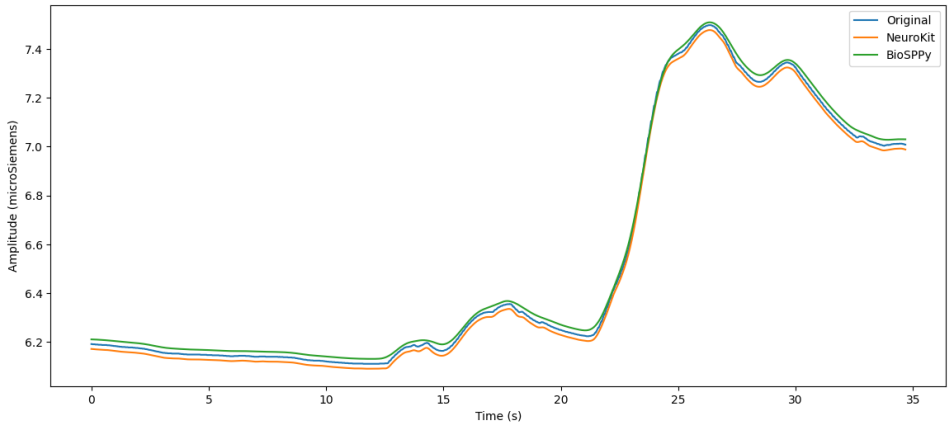


Figure 3.3: The raw signal and the two cleaned signals. Both methods use a Butterworth filter, but BioSPPy is more aggressive and also includes smoothing. The offset both signals have from the original signal is implemented manually.

EDA is generally split in a Tonic and a Phasic component. The Tonic component, also known as the Skin Conductance Level (SCL) represents the slow changes in the baseline. The Phasic component represents the fast responses, which are generally called Skin Conductance Responses (SCRs). Event-related SCRs generally arises 1-5 seconds after the event. To split these components different methods are used, In Posada-Quintero et al. 2020 different methods are discussed. There is no undisputed definition of which part is which component, and therefore the different methods can give quite varying outputs, as can also be seen in Figure 3.4. According to literature sparsEDA performs best for classifying event-related stress, also specifically in the context of drivers, through the Phasic component (Kumar et al. 2023; Lutin et al. 2021). However, a downside is a less accurate Tonic component (Lutin et al. 2021). CvxEDA (Greco et al. 2015) has a better tonic component and also performs well in the classification of the Phasic component. Since we do not just look at events, but want to correlate with continuous level of comfort, it was deemed important to preserve the tonic component.

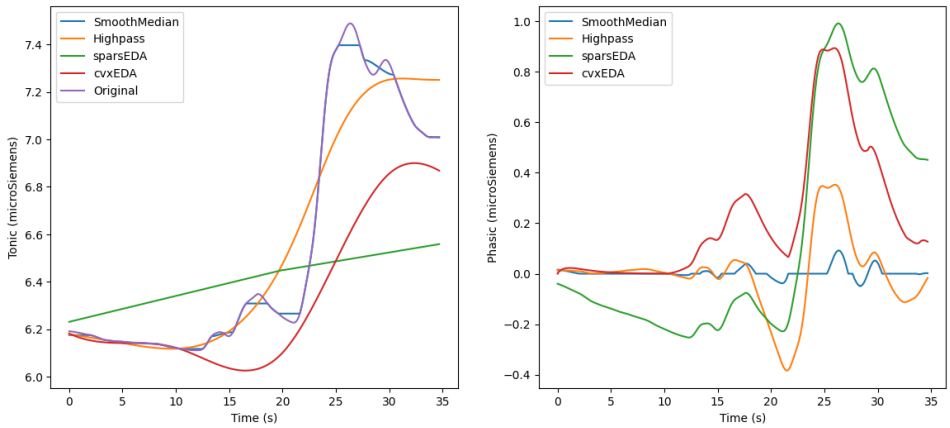


Figure 3.4: The Tonic (left) and Phasic (right) components of a signal, split through different methods.

In the Phasic component, peak detection can then be applied. For this, the popular method by Kim et al. 2004 was chosen, whom performed emotion classification from short physiological signals. These peaks represent the Skin Conductance Responses, and are expressed in three components. The Onset, where the rise starts, the amplitude at its peak, and the half-recovery time, when the level is halfway back to the base level. An overview summary of the processing is given in Figure 3.5.

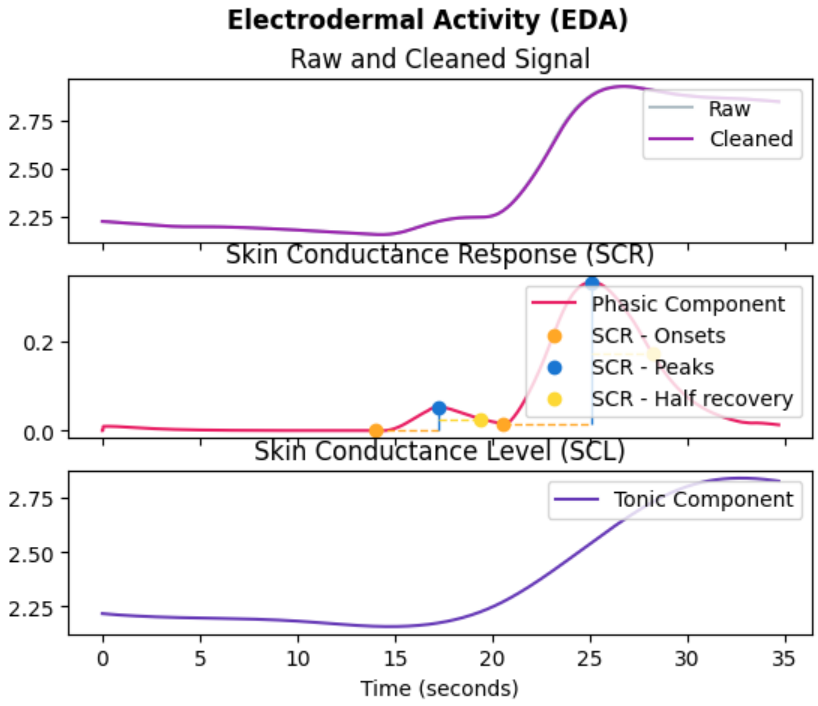


Figure 3.5: An overview of the processing of a GSR signal. First the signal is cleaned (top), then the Phasic (middle) and Tonic (bottem) components are extracted, and peaks are defined (middle).

3.4. Facial Emotion Recognition

In Chapter 2, the principle of facial emotion recognition was introduced, and the basics of neural networks were explained. In contrast to the physiological methods for driver state monitoring mentioned previously, facial emotion recognition can be a non-invasive method. It does not require the user to attach themselves to electrodes or wear specific devices. This makes facial emotion recognition an interesting area of research, not just in the context of automated driving, but also in other industries where user-state-monitoring is beneficial. For this research, existing models were specifically explored. In the end, the model by Zhang et al. 2023 was selected, due to its state-of-the-art performance, see also Figure 2.10, and availability of source code and pre-trained weights.

This model is trained on extracted faces from the dataset, which means using the model also requires to first extract the faces from the frames. To detect and extract the faces, RetinaFace was used (Deng et al. 2019). Retinaface is the same model as the developers of the DDAMFN++ model used for the pre-processing of their dataset Zhang et al. 2023. Retinaface is state-of-the-art for face detection. It detects 5 facial landmarks: eyes, nose, and the corners of the mouth. These landmarks are given in a dictionary with coordinates, in Figure 3.6 the detected facial landmarks are plotted, showing the model can detect these also for faces with glasses or facial hair.

Figure 3.6: Capture from the webcam during the experiment, with the left image showing the raw footage, and the right image showing the facial landmarks detected by RetinaFace.

After this detection, the face has to be cropped out of the frame. The DDAMFN++ model expects an input of size 112 by 112 pixels, which is also enforced by the pre-processing function. The facial area as detected by RetinaFace is not always square, and when it is not, it means the extracted face gets deformed a bit by the function. So instead of using the facial area that is given by RetinaFace, a different approach is used. Pre-defined coordinates are set, of where the facial landmarks should approximately be in the bounding box. Then a transfer function is generated that brings the facial landmarks as close to these coordinates as possible, without deforming the image. With this transfer function, the bounding box is used to extract the face from the image, which can be seen in Figure 3.7. The script written for this pre-processing is a python file that takes as input the path to the video file, and a path to the output folder where the extracted faces from each frame should be saved. This code can be found in Appendix A. Note that this can take some time, as the processing speed for this thesis was about 1.5 to 2 frames per second. RetinaFace was chosen for its high accuracy, however, when speed is more important, other available models like opencv are more lightweight.

Figure 3.7: Extracted face through the facial landmarks as depicted in Figure 3.6.

The architecture of the DDAMFN model is explained in more detail in the paper written by the developers (Zhang et al. 2023). To visualize the decisions that CNN-based models make, a technique called Gradient-weighted Class Activation Mapping (Grad-CAM) was applied. This is a technique to visualize the regions in an image that are important for the prediction the model makes by generating a heatmap Selvaraju et al. 2017. This heatmap is made by taking the gradients of the target class, which in this case is the predicted emotion, with respect to the features of the last convolutional layer. Each feature map is weighted, and by multiplying the feature maps with their weights a map is created highlighting the important areas. The heatmap for the picture we just extracted can be found in Figure 3.8. This figure shows the extracted face that is given as input, with the generated heatmap plotted over it. This shows that the model is well-focused on facial features.

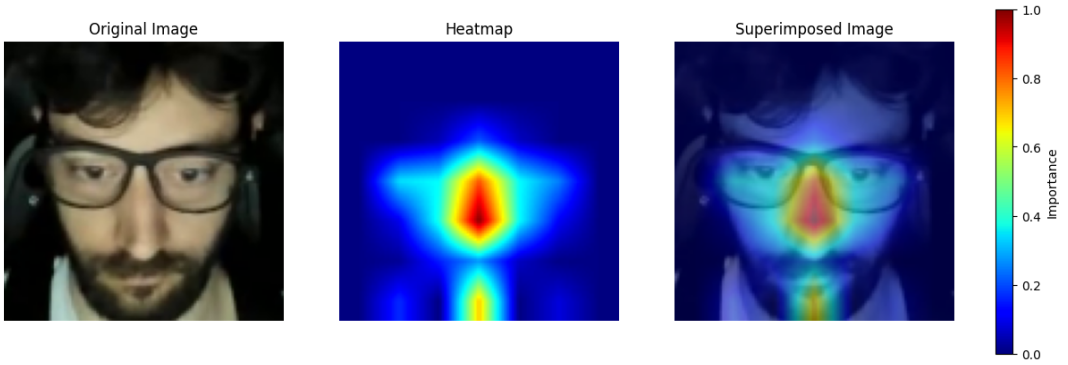


Figure 3.8: Heatmap showing the focus areas of the feature extraction. It shows that the model can focus on facial features.

3.5. Assessment model

In Chapter 2 some background in neural networks was given. For coding the assessment model, the PyTorch library in Python was used. The used model is based on the work of Mou, Zhou, et al. 2021. They use a CNN-LSTM-Attention network with multi-modal input data to predict the stress level of a driver. The three different categories of input data are physiological, vehicle dynamics, and environmental data. Even though the specific contents of these categories of data are different than the data used in this work, the similar goal of assessing the state of the driver through multi-modal data made it appealing to use this approach. The architecture of the model is summarised in Figure 3.9. A more detailed architecture including the specific parameters of each layer is given in Appendix G.

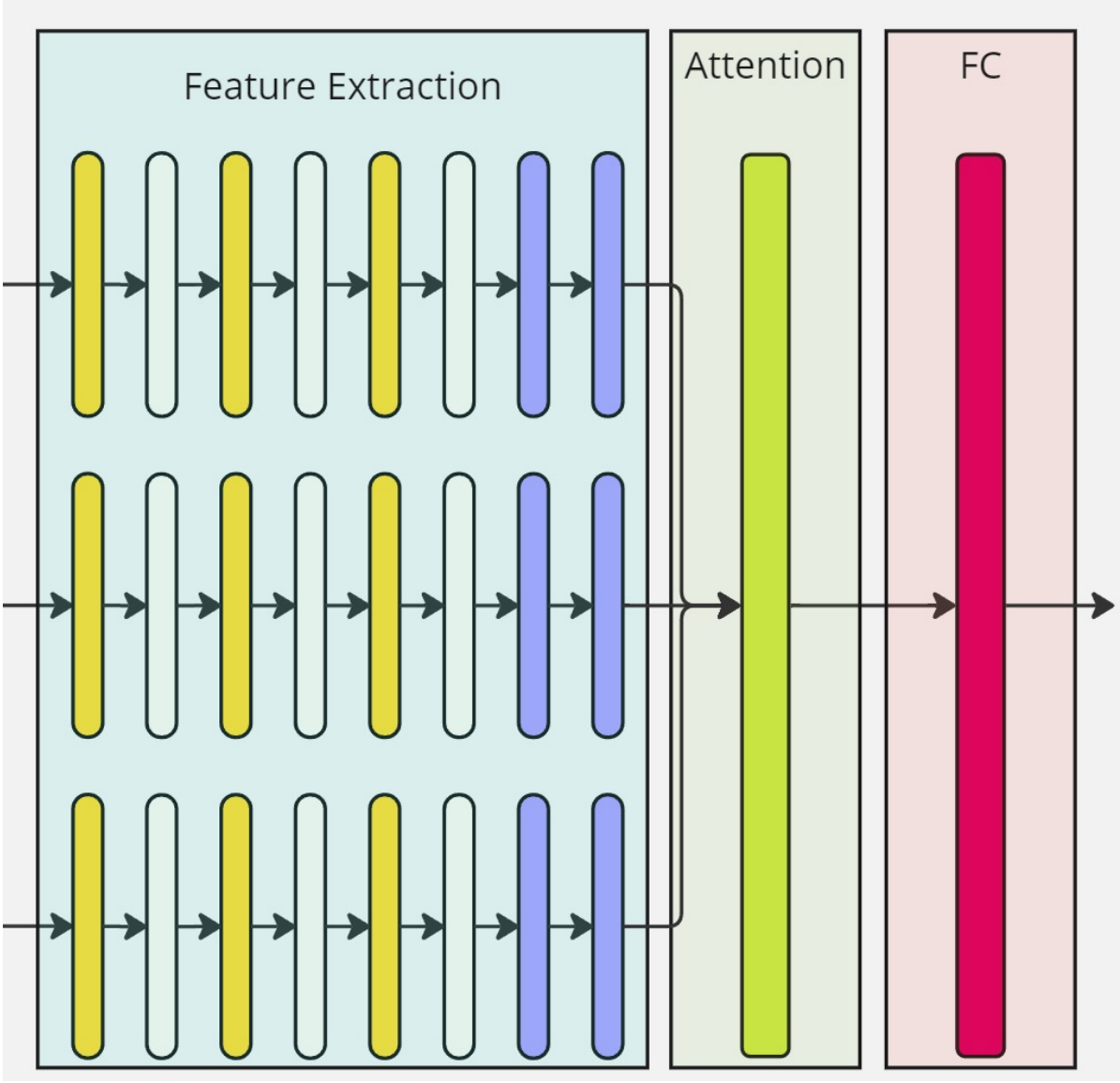


Figure 3.9: Model architecture. The three datatypes are given to their separate feature extraction model. In feature extraction, yellow represents a Convolutional layer, white represents Max-Pooling and Dropout, and blue represents an LSTM layer. The outputs from the three feature extraction models are concatenated and then attention is applied. The fully connected layer then gives the predicted comfort value as output.

4

Experiment

In this chapter, the experiment that was conducted is described. An elaborate description of the hardware that was used and the scenario itself allows for future research to replicate this experiment and compare results. It also provides a base for the discussion of the results, where recommendations will be made on how to approach this in future work. From the high-level overview presented in the introduction in Figure 1.1, this chapter zooms in on the part depicted in Figure 4.1. The experimental setup explained in this chapter consists of a driving simulator, a webcam, physiological sensors, and a comfort knob. During the experiment the participants experience a scenario in four different configurations, while their subjective scores and physiology is recorded. This data is analysed in Chapter 5.

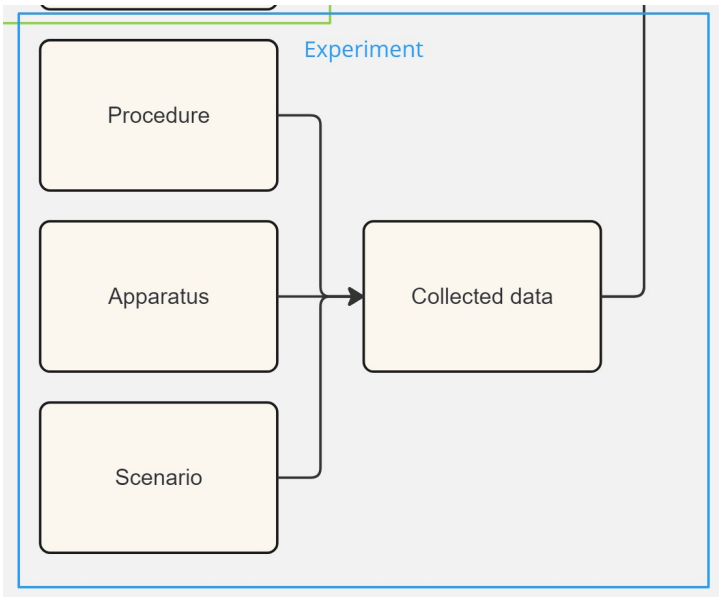


Figure 4.1: Overview of the Experiment chapter.

4.1. Apparatus

To best allow future research to replicate and/or bring improvements to this experiment, it is important to present the used equipment. Most of the equipment used in this experiment is commercially available, except for the comfort-knob which was developed in-house by Siemens, and the simulator platform, which was developed by MOOG.

4.1.1. Simulator

The driving simulator was the MOOG Hexapod, located in the Siemens facility in Leuven, with six degrees of freedom, which can be seen in Figure 4.3. On this platform a chair is placed which can be moved forward and backward with respect to the pedals, similar to a real vehicle. The scene is displayed on a wide-screen monitor of with a resolution of 3440 by 1440. The size of the screen is 34 inch / 86 centimeter. In Figure 4.10 we see this screen from the perspective of a participant, with the webcam in the middle above the screen, and on the bottom right there is a big red button that the participant could press should they want to stop the experiment. All of the participants were made aware of this option, though none of the participants used it or mentioned feeling the need to do so. On the operator table there was another one of these button, which also has not been used.

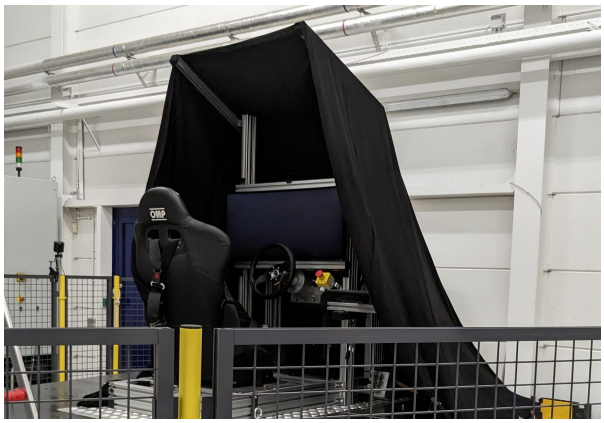


Figure 4.2: The simulator platform with the chair and the wide-screen.

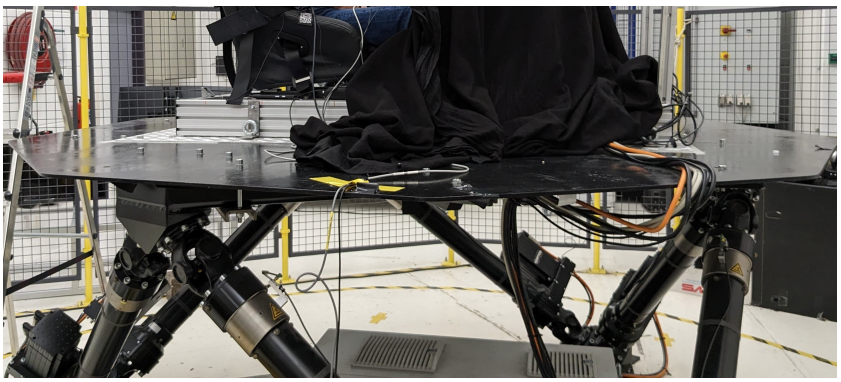


Figure 4.3: The MOOG Hexapod platform, with six degrees of freedom.

4.1.2. Sensors

During this experiment, measurements from five different sensors were recorded. A picture of someone with all sensors attached can be seen in Figure 4.4. For some of these sensors, an elaborate manual was written during the internship at Siemens. These manuals were mainly written for the next researcher or student at Siemens who would use the devices, and can be found in Appendix E. A picture of one of the participants with all devices setup can be found here in Figure 4.4.

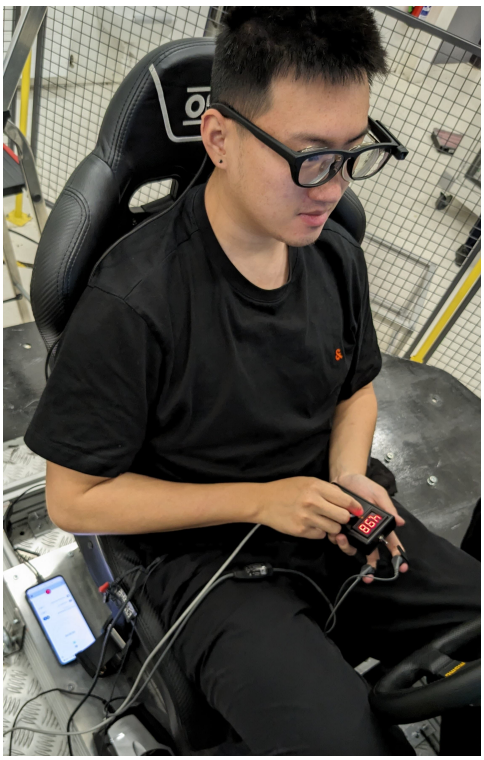


Figure 4.4: One of the participants in full setup. In the picture the eye tracking glasses are on his head, connected to the phone that can be seen in the bottom left. The phone is held in place using velcro tape. In his hand, the participant is operating the comfort knob. on his left index and middle finger, the velcro straps with GSR electrodes are attached. The GSR electrodes and heart rate electrodes, which are under the participants' shirt, are connected to the grey/black NeXus device next to the phone. This device is also held in place using velcro tape.

Comfort-Knob

The Comfort-Knob is a rotational potentiometer connected to a Simcenter SCADAS system to continuously measure the subjective comfort of the participant. It is a turning knob where the participant gives a value between 0 and 10. The scale was predefined as 0 being extremely uncomfortable, 5 being neutral, and 10 being extremely comfortable. This is based on the SAE J1060 for subjective ratings related to ride comfort in motor vehicles, where 0 up to and including 4 is considered unacceptable, 5 is border line, and 6 to 10 is acceptable (Cieslak et al. 2020). Except for the knob not turning past the limits of 0 and 10, there was no haptic feedback for the participant to feel how far they had turned the knob. Figure 4.5 was constantly present in the top-middle of the screen to remind the participant which way to turn the knob according to their state. This illustration was made very simple on purpose, to not for a distraction or leave things open for interpretation. On the device itself is a LED screen that also shows the current value in 2 decimals. They are told not to focus too much on the exact value, as this would make the activity of giving the ratings more of a distraction. A picture

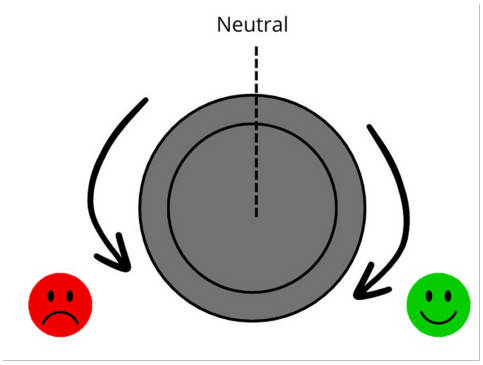


Figure 4.5: The illustration that was in the top middle of the screen for the participants, to remind the participant which way to turn the knob.

of the comfort knob is presented in Figure 4.6.

Figure 4.6: The Comfort Knob. A marking was made on the neutral position, and a figure was placed in the screen to help the participant remember which way to turn the knob.

Heart rate & Galvanic Skin Response

The aim of this experiment was to collect a dataset as complete as possible, which is why even though heart rate is not analysed in this thesis, it was still collected for future research.

Heart rate and skin conductance were measured using the NeXus-10 MKII device from MindMedia. The full manual for this can be found in Appendix E. The heart rate is measured via three electrodes placed on the torso of the participant. The electrodes are color-coded and placed as illustrated in Figure 4.7. There are multiple electrode stickers available on the market. The specific sticky electrodes used in this experiment were the F9060 electrodes (for adults, 48x50 millimeter) produced by FIAB. The participant attached these stickers themselves, under clear instruction from the researcher.

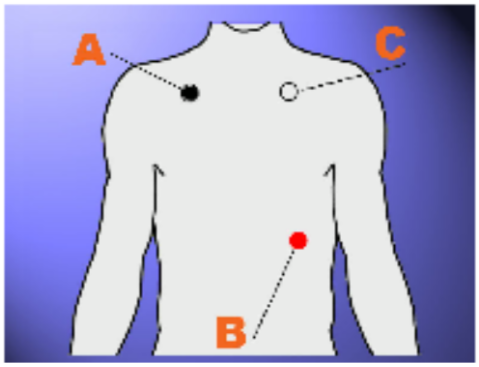


Figure 4.7: Illustration on where to place the electrodes for heart rate measurement. This is the front view of the torso, so from the perspective of the participant: A is located top-right, B bottom-left, and C top-left. A (black electrode) is the negative input, and B (the red electrode) is the positive input. C (the white electrode) is the ground electrode.

The Galvanic Skin Response is measured via two electrodes attached to two fingertips of the same hand with velcro straps, as can be seen in Figure 4.8. It does not matter whether this is the right or left hand. For this experiment, the participant was asked what was their dominant hand, and how they would operate the comfort knob. The velcro straps were attached to the fingertips on the non-dominant hand or the hand that would not operate the comfort knob. This was done to avoid excessive movements of the electrode, as that could affect the measurements. With the GSR electrodes, it is important that there is some time (1 or 2 minutes) between attaching the sensor and starting the measurement, as the skin under and around the velcro straps has to accommodate to the presence of the straps. During this experiment, after the velcro straps were attached, explanations on the other sensors were given, and a test run was done, ensuring enough time before starting the measurements.

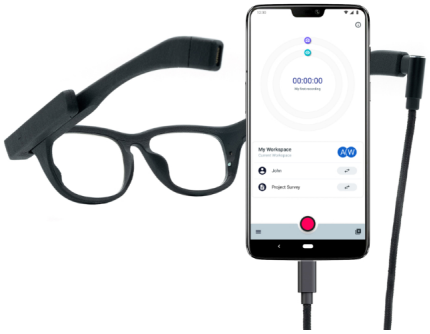
Figure 4.8: The velcro straps with GSR sensors attached to the fingertips of the non-dominant hand. They are not color-coded because it does not matter which goes on which finger, as long as they are on the same hand.

Eye-tracking

The aim of this experiment was to collect a dataset as complete as possible, which is why even though eye-tracking is not analysed in this thesis, it was still collected for future research.

Eye-tracking is done with the Pupil Invisible glasses by pupil-labs. The user manual written for this device is in appendix E. Note that from this manual the login codes are retracted, as they are for the Siemens device and account specifically. These glasses contain a 6 degrees of freedom IMU, a scene camera for first person view which also includes a microphone to capture audio, and two infrared eye cameras to capture eye videos and perform real-time gaze estimation. The glasses are connected via USB-C to a smartphone that has the Invisible Companion application installed. Via the interface of this application measurements are started and stopped, see also Figure 4.9b.

(a) Locations of the hardware present in the Pupil Invisible. For sensors it contains an 6 degrees of freedom IMU, a scene camera for first-person view, and an infrared eye camera for to capture eye videos, and real-time gaze estimation.



(b) The Pupil Invisible glasses come with a smartphone with their own software application on it. Measurements are set-up and started from this phone.

Figure 4.9: The Pupil Invisible developed by Pupil-Labs. These images are from their technical overview, where they also have elaborate documentation and expelenations.

Webcam

The face is recorded via a webcam located above the screen, see also Figure 4.10. The frames are timestamped in Python with UNIX timestamps in nanosecond precision. These timestamps are stored in a data frame together with the frame number. The frames are in color and saved as a video of 30 frames per second with a resolution of 640 by 480. The script that was used to save the images can be found in Appendix A.

Figure 4.10: First person view from the participants. The webcam is located in the top middle above the screen.

4.2. Scenario

The scene for this experiment was a so-called ”real2sim” conversion of a vehicle drive performed by Siemens in a previous experiment. During this experiment, a real vehicle was driven around a track, in different driving styles. Data collected included all the above-mentioned sensors, except for webcam footage. The front-view camera of the vehicle was used, and from this footage, the Simcenter Prescan scenario was created. The perspective of this camera can be seen in Figure 4.11, this is the scene as the participants experienced it during the experiment. The scenario starts with a straight line, then a wide right-hand corner, right after which a stop sign is placed where a vehicle facing the other direction. After that, it is again a straight line. A birds-eye view of the experiment with pointers and indicators for the different objects and roads can be found in Figure 4.12. In the experiment, the participants experienced this scenario in 4 different setups. There are two driving styles, one being more dynamic and one being more calm. These driving styles are from the real-vehicle experiment, where the same driver consistently drove the vehicle. Both these styles are experienced with and without a pedestrian crossing at the stop sign. The view of the pedestrian crossing the road from the perspective of the participant can be seen in Figure 4.11c. In terms of dynamics, the only difference between a scenario with and without the pedestrian is the stationary period. Other than that the dynamics for those scenarios are the same. This can be seen in the acceleration profiles in Figure 4.13. In Figure 4.14 the time-to-collision (TTC) is plotted for both driving styles. Here it is visible that the dynamic scenario is more dangerous, with the TTC decreasing faster and coming to a lower value before the vehicle stops.

(a) View of the starting position of the scene.

(b) View of the beginning of the right-hand corner.

(c) View of the pedestrian crossing.

Figure 4.11: Screen captures from the scene. This is the perspective from which the participants experience the scene.

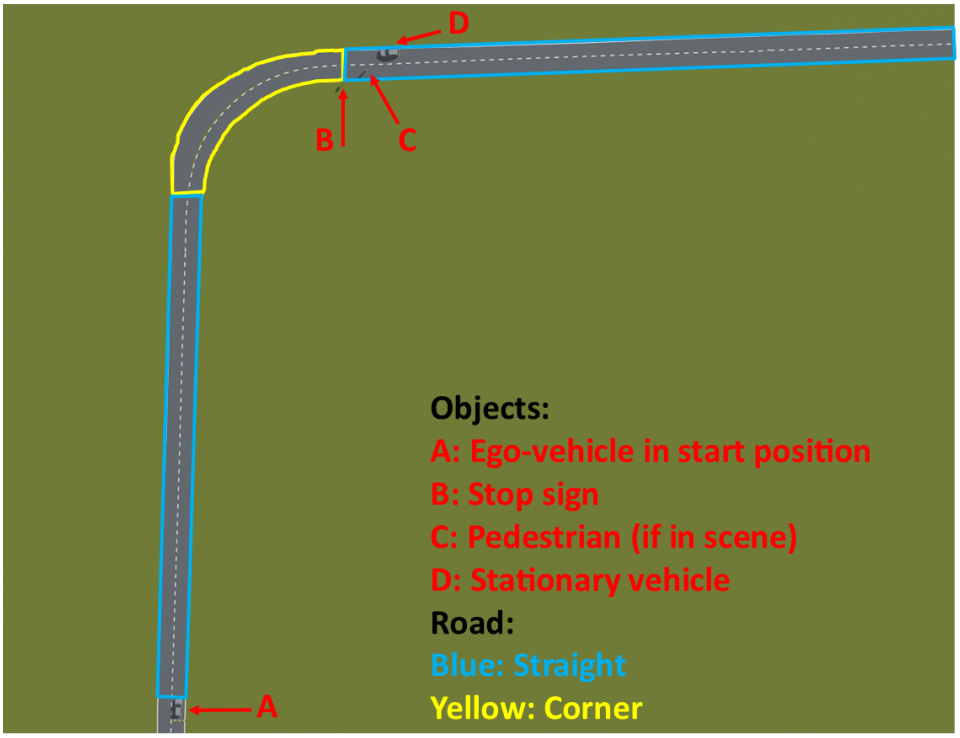


Figure 4.12: Birds-eye view of the scene, with the different road types, straight line, and right-hand turn, indicated with blue and yellow boxes respectively. The objects in the scene are indicated with red arrows and a letter: A is the ego vehicle in the starting position, B is the stop sign, C is the pedestrian, and D is the stationary vehicle. Note that the pedestrian is not in the scene for every configuration.

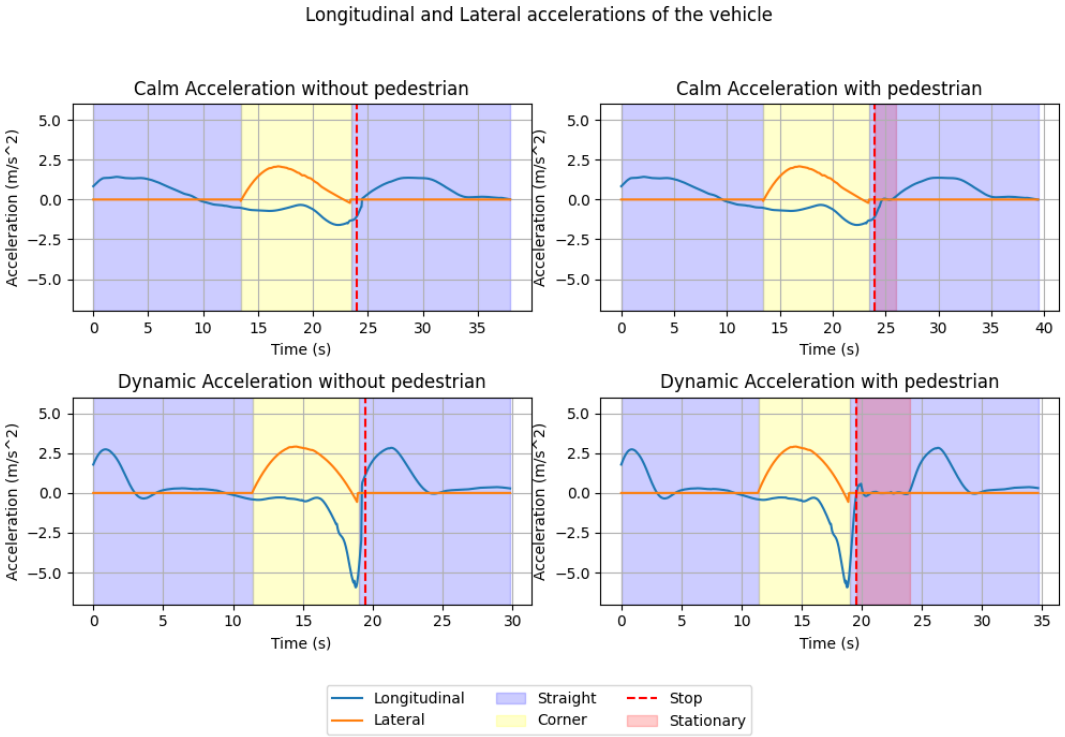


Figure 4.13: Linear Acceleration profiles of the four different configurations. The colors are the same as on the birds-eye view in Figure 4.12. Blue is the straight line, yellow is the right-hand turn, and the red striped line marks the point where the vehicle stops. If the pedestrian is in the scene, the red part is the stationary time for the pedestrian to cross.

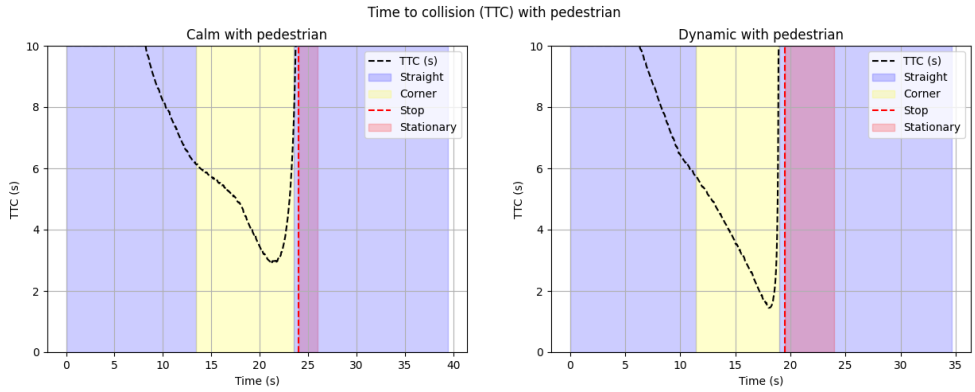


Figure 4.14: Time-to-collision with the pedestrian for both driving styles. Both axes are in seconds, with the x-axis representing the time point in the run, and the y-axis the TTC. The colors are the same as on the birds-eye view in Figure 4.12. Blue is the straight line, yellow is the right-hand turn, and the red striped line marks the point where the vehicle stops. If the pedestrian is in the scene, the red part is the stationary time for the pedestrian to cross.

4.2.1. Vehicle dynamics data processing

The simulator takes the following data as input, as a .mat file, over time in a frequency of 1kHz:

- Linear accelerations (x, y and z direction - [m/(s^2)])
- Angular accelerations (Roll, Pitch, and Yaw - [rad/(s^2)])
- Angular velocities (Roll, Pitch, and Yaw - [rad/s])
- Forward velocity ([m/s])
- Angular position (Roll and Pitch - [rad])
- Prescan data (for any dynamic object in the scene):
  - Position (x, y, and z - [m])
  - Orientation (Roll, Pitch, and Yaw - [rad])

The linear accelerations and angular velocities can be obtained directly from the Septentrio IMU that was installed in the vehicle. This collected data at 20Hz, however, after reviewing the data it could be seen that the time interval was not consistent. Some time steps were missing, while other points seemed to have a double measurement, with two timestamps extremely close together. A constant timestamp was desired for further processing, so Piecewise Cubic Hermite Interpolating Polynomial interpolation from the SciPy library in Python was used to resample the data. This resamples the data to constant time-steps of 0,05 seconds, with the given constraints of the first and last time-step, while preserving the monotonicity of the data and not overshoot, even when the data is not smooth (Fritsch et al. 1984).

The vehicle simulator works with the transform as defined by the SAEJ670 (Committee 2022): The positive x points in the forward direction, the positive y points to the right, and the positive z points down. The data from the vehicle were given with different coordinate frames. For the linear accelerations, they are the exact opposite, with x pointing backward, y pointing left, and z pointing up. The angular velocities were given with x pointing forward, y pointing right, and z up. For a clarification of the desired transform, see Figure 4.15.

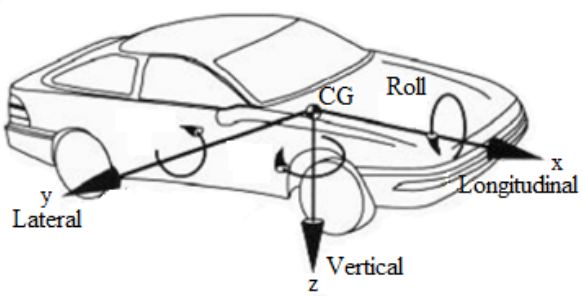


Figure 4.15: The axis system of the vehicle simulator. The angular directions follow the right-hand rule with Roll around X, pointing forward, Pitch around the Y, pointing right, and Yaw around the Z, pointing down. This system follows the definition of SAE J 670 (Committee 2022; Macfarlane 2016).

The velocity of the vehicle was measured by the GPS device. However, this was not in a local coordinate frame, but in a global GPS frame. Getting the transform for this would be difficult as it depended on the orientation of the vehicle. After discussing this issue with the perception expert and simulator expert of Siemens, it was concluded that since there was no backward driving at any moment in the scenario, and the simulator uses this value as speed relative to the ground for the motion queuing, using the absolute value for the velocity would not be an issue.

The directional and angular positions came from a different device, giving x, y, and z positions in the GNSS, the coordinate system that was also used to build the road network in Simcenter Prescan. The orientation is given via quaternions, which were converted to the Euler angles using the following formulas (Blanco-Claraco 2021):

$$\begin{bmatrix} \phi \\ \theta \\ \psi \end{bmatrix} = \begin{bmatrix} \text{atan2}\left(2(q_w q_x + q_y q_z), 1 - 2(q_x^2 + q_y^2)\right) \\ \text{asin}\left(2(q_w q_y - q_x q_z)\right) \\ \text{atan2}\left(2(q_w q_z + q_x q_y), 1 - 2(q_y^2 + q_z^2)\right) \end{bmatrix} \tag{4.1}$$

As this was another device, it did not follow the same timestamps, which were in UNIX format and nanosecond precision for both. To align this, a function was made to align the closest timestamp of the GNSS device with every IMU timestamp. As the final step, the full dataset was interpolated via linear interpolation to 1kHz, converted to a .mat file, and uploaded to the workstation of the simulator.

4.3. Procedure

Before starting, the participant was given a printed consent form to read and sign. This consent form was set up in collaboration with and approved by Jakob Oczko, data privacy manager at Siemens, and can be found in Appendix C.

Then the participant was given a tablet with a keyboard to fill in a questionnaire via Google Forms. This questionnaire is used to collect some demographic data and can be used to assess subjective pre-existing trust and potential change in trust in automated vehicles. The questionnaire can be found in Appendix B.

After this, the participant was guided to their seat in the simulator, where the different sensors were explained and attached. Especially for the GSR sensor it is important that there is some time between attachment and the start of the experiment, to let the body accommodate to its presence. When all sensors were attached and possible questions of the participants were answered, a test run of a scenario without a pedestrian was done. This had the purpose of letting the participant get familiar with the scenario, the simulator itself, and the comfort-knob. For the comfort-knob, it was emphasized to the participant that they should not focus too much on the exact value displayed. This was to avoid participants looking down at the knob all the time and have them focus on the driving experience.

After the test run, the curtain was closed around the participants in preparation for the experiment to start. Before closing the curtain the participants were asked if this was alright with them. None of the participants mentioned suffering from e.g. claustrophobia and all were fine with the curtain closed around them. It was explained to them that they would experience the same scenario in two different driving styles, without knowing in which order beforehand. After the first two runs, they were asked to identify which was the calm driving style and which was the dynamic driving style. All but one participant was able to identify these first two runs correctly. Then they were told that these two driving styles would

be repeated, and the order could be either the same or different. They were not told, that now in both driving styles, there would be a pedestrian crossing the road. This added effect of surprise was intended to stimulate reactions in the physiological data. With the pedestrian in the scene, all participants could correctly identify which order the driving styles were presented.

Then all four different configurations (calm with pedestrian, calm without pedestrian, dynamic with pedestrian, dynamic without pedestrian) were repeated in random order, where the participant was asked to identify which it was after each run. During this session, a total of 8 out of 128 runs were classified incorrect, spread over 7 participants. All of these were for scenarios without the pedestrian in it. Before starting these four runs, the participant was told that there would be one difference, namely that now in one of the runs it was possible that they would hit the pedestrian. This was not true, the presented experiments were identical, however, it had the purpose of keeping them more engaged with the scene, because experiencing the same scenario multiple times on the simulator, and thus already knowing what will happen, could make it boring. It is certain that not every participant completely believed this, as some were already familiar with the simulator or with Simcenter Prescan, but even if the participant would only slightly consider the possibility it would serve its purpose. However, one participant mentioned that they thought they should not have been told about this possibility, as now the participant was actually ”warned” about the dynamic driving style being more aggressive. When the participant was told that it was the same run as they had experienced before, the reaction was surprised as the experience was that it was more aggressive.

After the last run was done, the sensors were stopped and the data was saved. The participant was assisted in removing the sensors and getting out of the simulator, and asked how they felt about the experience. Afterwards, they were given the tablet again with another google form containing the same questions on trust and comfort in automated vehicles as before the experiment, to see if something changed. They were also asked to elaborate on their answers in case something had changed.

4.4. Collected Data

The following data were collected during the experiment:

- **Subjective Comfort:**
  - Continuous comfort rating: Participants used a comfort knob to provide a continuous rating from 0 (very uncomfortable) to 10 (very comfortable).
  - Questionnaire responses: Participants answered questions regarding their initial trust and any changes in trust due to the experiment.
- **Simulator Dynamics:**
  - Positions: Spatial coordinates of the simulator.
  - Orientations: Angular orientations of the simulator.
  - Angular velocities: Rate of change of orientation.
  - Linear velocities: Rate of change of position.
  - Angular accelerations: Rate of change of angular velocity.
  - Linear accelerations: Rate of change of linear velocity.
  - Longitudinal accelerations: Forward and backward accelerations.
- **Biometric Data:**
  - Galvanic Skin Response (GSR): Measures skin conductance.
  - Heart rate: Number of heartbeats per minute.
  - Heart rate variability (HRV): Variations in the time interval between heartbeats.
    - Amplitude: Magnitude of HRV.
    - Low-frequency power (0.04-0.16Hz): Power in the low-frequency band of HRV.
    - High-frequency power (0.16-0.4Hz): Power in the high-frequency band of HRV.
    - LF/HF ratio: Ratio of low-frequency to high-frequency power.
- **Webcam Footage:** Video footage of participants’ faces was recorded using a webcam, capturing facial expressions during the experiment.

5

Data analyses and Results

In this chapter part of the collected data from the experiment as described in Chapter 4 is analysed, using some of the methods that are described in Chapter 3. From the high-level overview presented in the introduction in Figure 1.1, this chapter zooms in on the part depicted in Figure 5.1. First the subjective comfort ratings have been analysed, confirming that the experiment was successful in eliciting different levels of comfort. Then the results from facial expression recognition model are analysed, attempting to answer the main research question. Finally, a comfort prediction model was set up using vehicle dynamics and GSR data. Due to the amount of data collected, and this only being one master thesis project, not all signals were analysed. The dataset by itself still provides opportunities for future research.

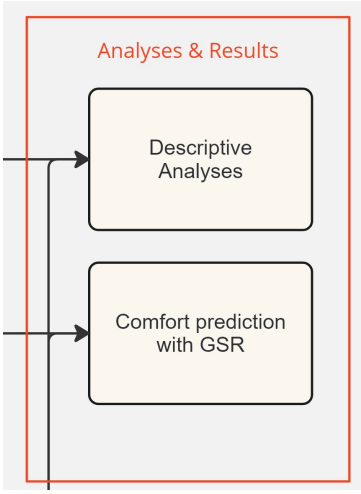


Figure 5.1: Overview of the Data analyses and Results chapter.

5.1. Descriptive analyses

Overall 36 participants took part in the experiment. The first 4 were considered pilots, as the experiment was still being evaluated and improved. Thus, their experience was not exactly the same as that of the rest of the participants and their data was not included in the analyses. Participant 5 through 36 all experienced the full experiment in its finalised setup. Many said they enjoyed the experience, and only two participants mentioned some unease that could be classified as minor motion sickness, but they did not call it that themselves and no symptoms that would be classified higher than 1 on the MIsery SCale (MISC). We can conclude that motion sickness did not have an impact on the results.

5.1.1. Continuous subjective comfort ratings

The participants experienced four different configurations of the driving scenario, presented in Table 5.1. First, we report the subjective comfort ratings. All scenarios were experienced two times by the participants, the first time in a set order: Both driving styles were first played without a pedestrian, and then repeated without telling the participant that there would be a pedestrian in the scene. After the ”set” session, all scenarios were repeated in random order, giving a total of 8 runs per participant. All results are plotted in the figures on the following pages (Figure 5.2, 5.3, 5.4, and 5.5). Each page contains all plots for one configuration. On the x-axis is time in seconds, and on the y-axis is the comfort score between 0 and 10. The blue line shows the score given during the first round, and the red line is the second round.

Table 5.1: The four different configurations that the participants experienced.

Driving Style	Pedestrian in the scene
calm	no
calm	yes
dynamic	no
dynamic	yes

First, we have to analyze if all the data is good, or if some need to be excluded. The data for Participant 20 during the second round for the calm driving style without the pedestrian was unfortunately lost, which is why this plot also has no red line. Participant 11 misunderstood the working of the knob during the first round, meaning this knob data is also not viable. All data for participant 19 was also excluded from further analyses, as they gave unrealistic scores and after the experiment also commented that it was normal for them if a car also ”nudged” a pedestrian. This was considered very exceptional, and not relevant to how we desire automated vehicles to behave, which is the direction this area of research eventually aims for. Participants 31 and 32 seem to only adjust the comfort knob at the very end of the run, long after the critical scenario presented itself, or not at all, leading to believe that they might have misunderstood the assignment, or forgot to use the knob till the very end. Their knob data was also excluded.

The main purpose of repeating these scenarios is to check for consistency. If the subjective score is consistent over multiple runs, it is more likely to show a true representation of the participants’ state. This continuity was assessed visually by looking at these graphs, and also a dynamic time-warping (DTW) function was applied starting at 10 seconds, to account for initial offsets, to see if this aligns with the observations.

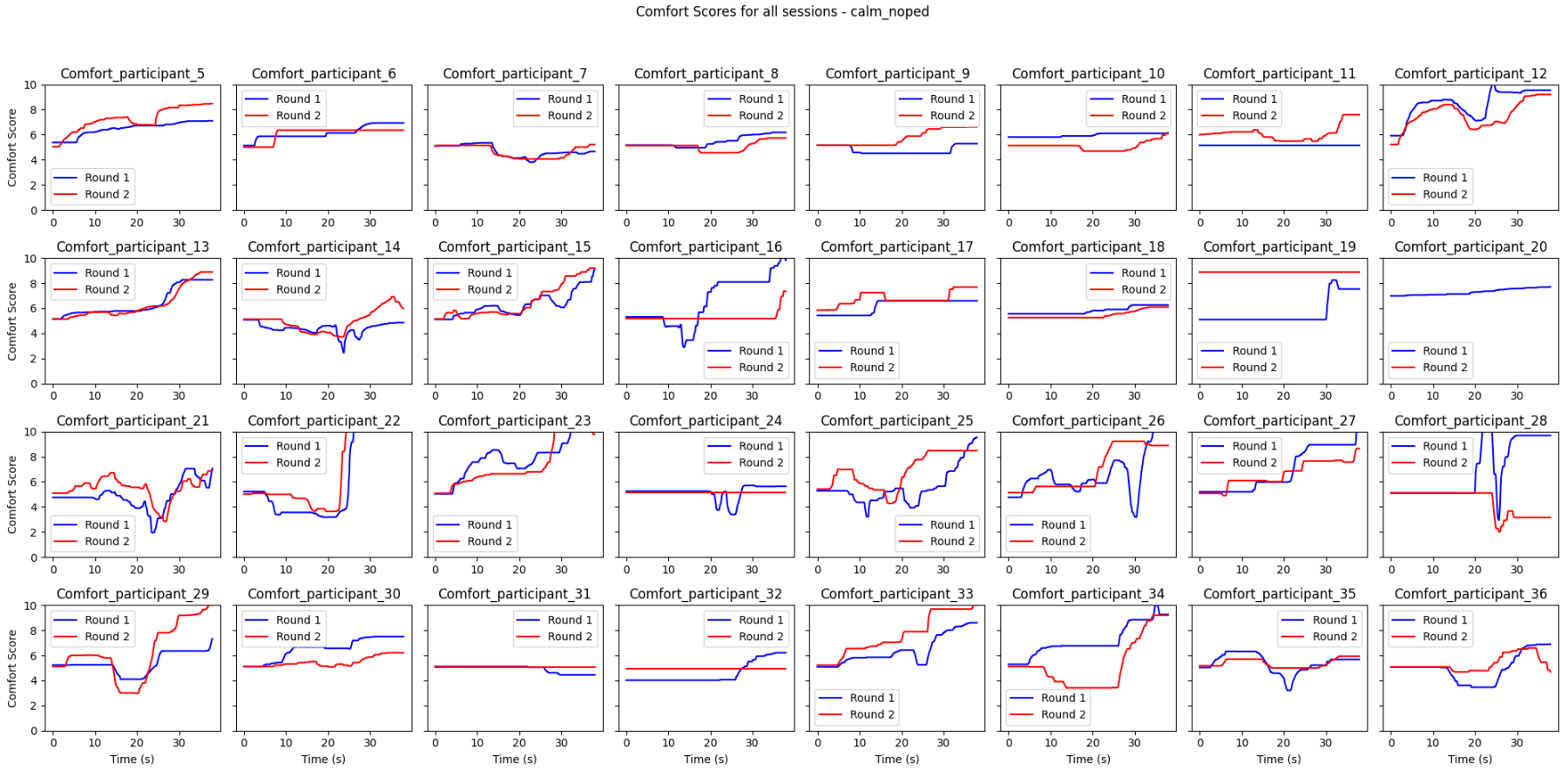


Figure 5.2: All subjective comfort responses for the calm drive without the pedestrian in the scene. Each blue line represents the response during the first round, and the red line the response during the second round.

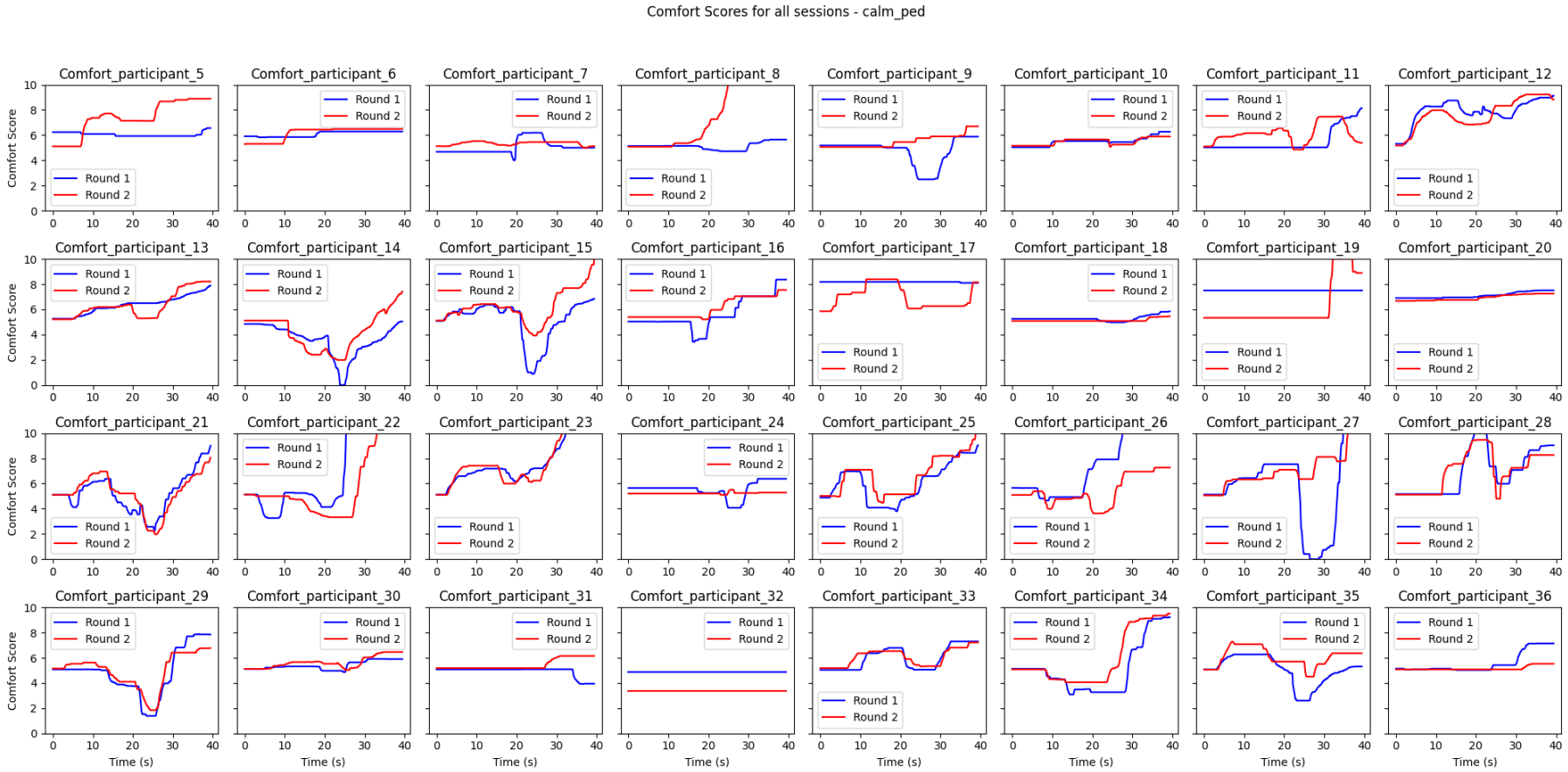


Figure 5.3: All subjective comfort responses for the calm drive with the pedestrian in the scene. Each blue line represents the response during the first round, and the red line the response during the second round.

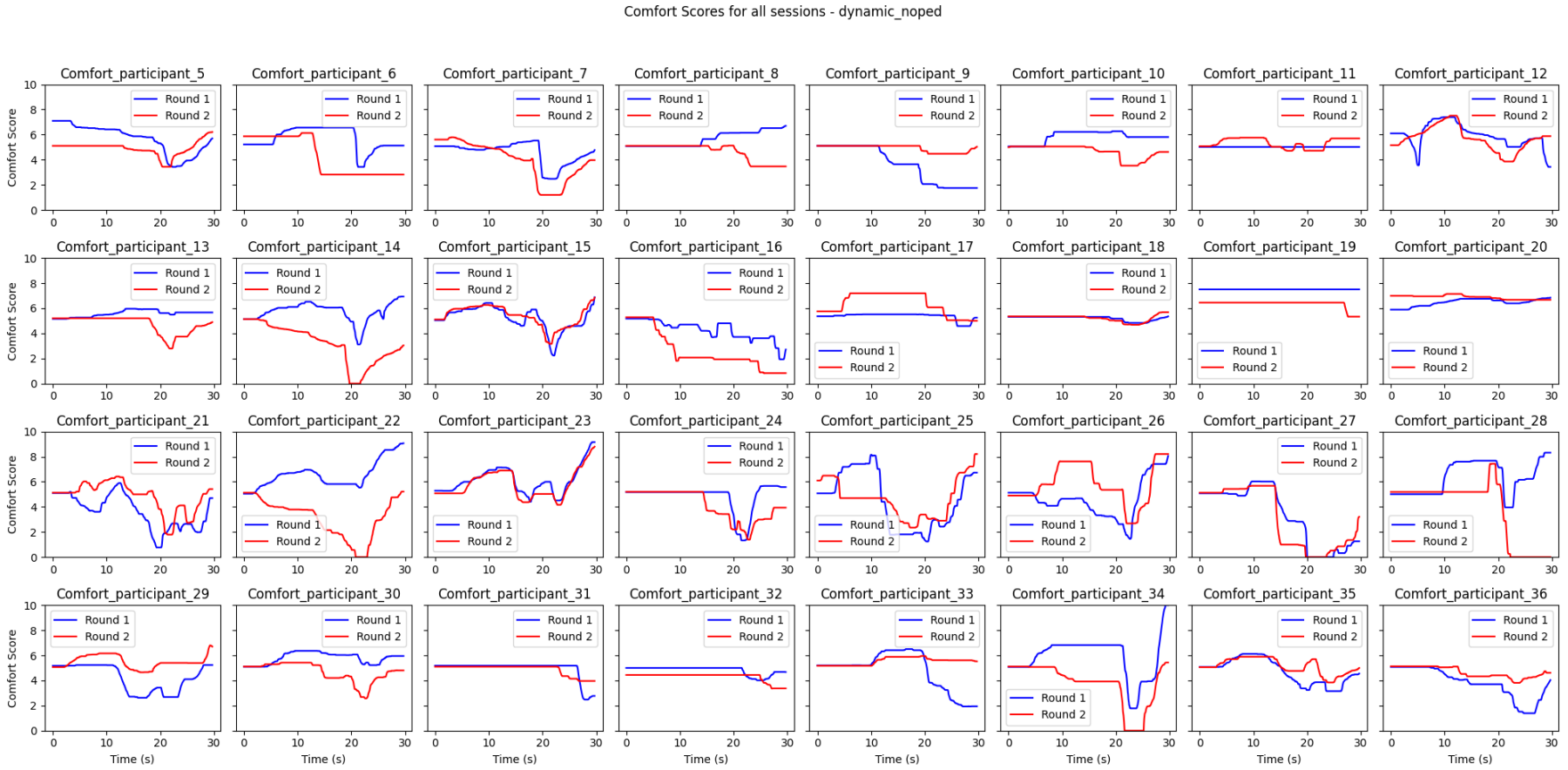


Figure 5.4: All subjective comfort responses for the dynamic drive without the pedestrian in the scene. Each blue line represents the response during the first round, and the red line the response during the second round.

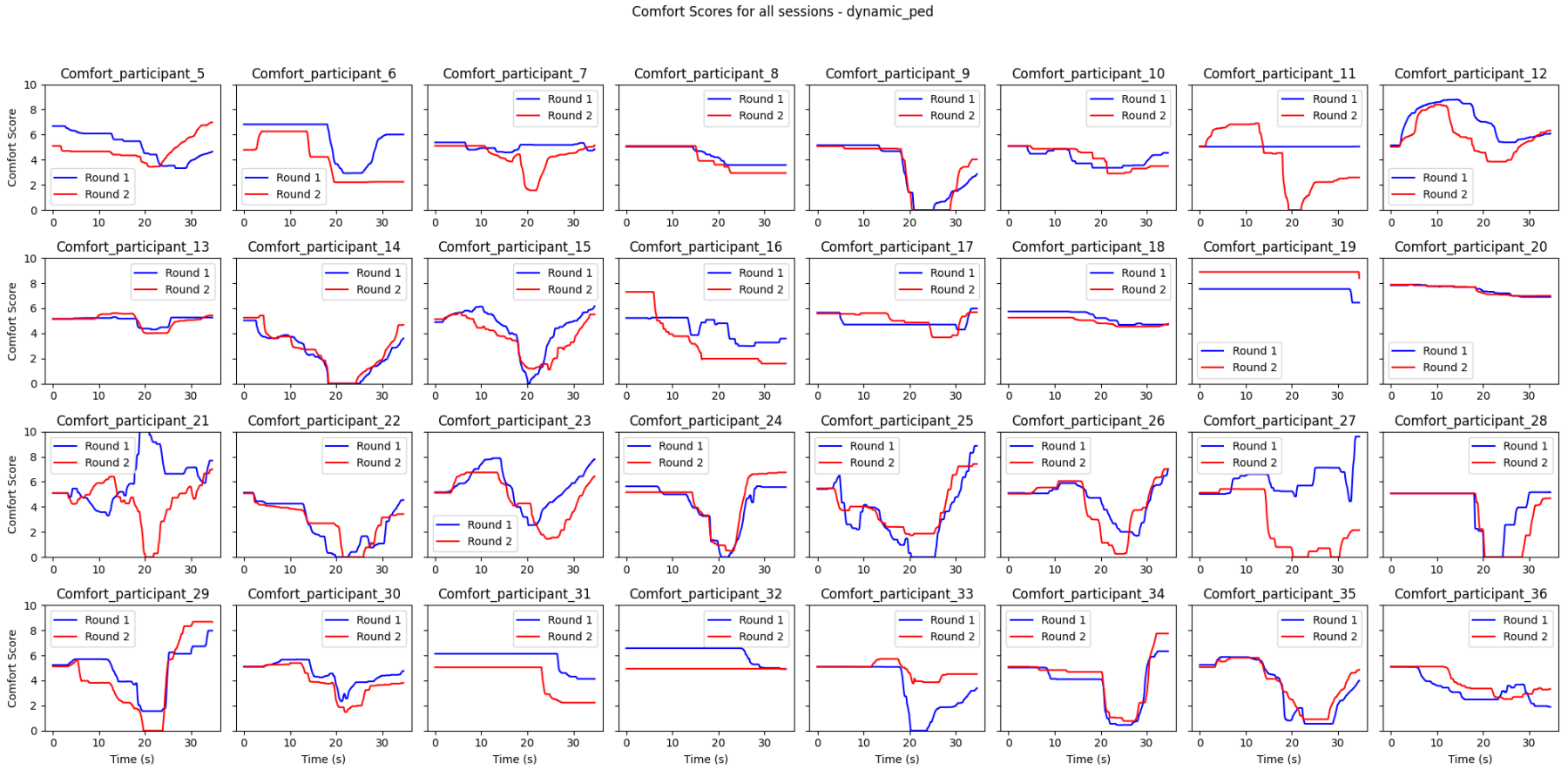


Figure 5.5: All subjective comfort responses for the dynamic drive with the pedestrian in the scene. Each blue line represents the response during the first round, and the red line the response during the second round.

The resulting DTW distances are presented in Table 5.2. The higher the distance, the less similar the two plots for that configuration are according to the DTW method. Participant 21 seems to have given the exact opposite rating during the first round of the dynamic drive with the pedestrian. Looking at their consistency in the other configurations, it is likely that they inverted the comfort scale during that run.

Table 5.2: DTW distances per participant per configuration. The higher the value, the less similar the two subjective ratings given by the participant for that round are. The excluded participants are made bold

	calm_noped	**calm_ped**	**dynamic_noped**	**dynamic_ped**
Continuity participant 5	30,21	**66,97**	18,85	34,83
Continuity participant 6	15,01	6,13	38,7	61,53
Continuity participant 7	6,47	18,31	19,56	**37,47**
Continuity participant 8	10,85	**110,08**	**53,61**	14,19
Continuity participant 9	35,05	**42,33**	**56,95**	10,76
Continuity participant 10	28,04	5,67	46	13,69
Continuity participant 12	13,35	12,9	20,21	25,26
Continuity participant 13	8,07	11,88	38,52	6,72
Continuity participant 14	28,55	29,18	**75,99**	7,71
Continuity participant 15	9,26	43,7	8,79	15,41
Continuity participant 16	**38,66**	22,5	**57,96**	41,51
Continuity participant 17	24,86	48,38	34,92	16,83
Continuity participant 18	7,83	5,6	3,29	8,67
Continuity participant 20		5,36	7,5	1,45
Continuity participant 21	21,04	9,9	15,42	**62,89**
Continuity participant 22	19,17	15,54	**115,36**	9,83
Continuity participant 23	26,09	5,69	9,72	20,93
Continuity participant 24	21,44	24,93	25,01	19,93
Continuity participant 25	21,72	9,73	28,6	28,61
Continuity participant 26	32,78	**64,16**	46,8	17,05
Continuity participant 27	14,08	**108,59**	8,77	**120,66**
Continuity participant 28	**143,46**	13,81	**120,31**	6,5
Continuity participant 29	43,42	17,53	40,05	30,64
Continuity participant 30	36,73	11,42	38,85	12,88
Continuity participant 33	30,17	8,46	**56,27**	**65,94**
Continuity participant 34	**81,51**	18,21	**63,86**	19,76
Continuity participant 35	18,58	34,63	11,13	9,92
Continuity participant 36	29,04	28,24	33,62	23,56

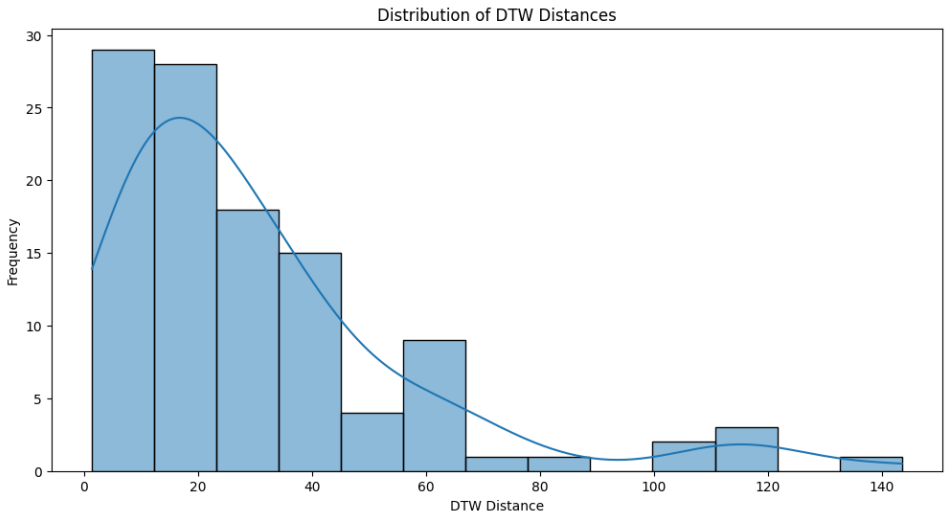


Figure 5.6: The distribution of the DTW distances. This gives an idea of what continuity can reasonably be expected from the participants.

The distribution of the DTW analyses is shown in Figure 5.6. This gives an overview and a general idea of what can be reasonably expected from the participants. A soft cut-off was selected, at the 85th percentile, corresponding to a value of 54.94. This was then combined with some visual inspection. Some exceptions were made, and explained here:

- **Participant 16, calm noped:** A score of 38.88 but round 2 barely changes and round 1 a lot. the long straights make it seem similar, but the absolute difference between the two ratings is big for most of the run.
- **Participant 9, calm ped:** A score of 42,37. The score is not too low only due to the exact overlap up to almost 20 seconds into the run, however, after that the rounds diverge, feeling discomfort in round 1 and slight comfort in round 2.
- **Participant 8, dynamic noped:** A score of 53.61, almost the cut-off. The moment the person changes their comfort, they do the exact opposite in either round. It is possible that they inverted the scale mistakenly in one of the rounds.
- **Participant 6, dynamic ped:** A score of 61.53, mainly because at the end they set their comfort back up in round 1, and not in round 2. However, before that during the scenario the trends are similar, so it was decided to keep this in.
- **participant 7, dynamic ped:** A score of 37.47. Again, barely changing comfort in round 1, but in round 2 reporting strong discomfort around 20 seconds.

The final selection of participants whose comfort score is excluded from further processing is in Table 5.3.

Table 5.3: Excluded **comfort knob data** for participants per configuration based on DTW and visual inspection of the plots. *for this configuration, only the first round of participant 21 is removed.

Configuration	Excluded Participants
calm_noped	11, 16, 19, 28, 31, 32, 34
calm_ped	5, 8, 9, 11, 19, 27, 31, 32
dynamic_noped	8, 9, 11, 14, 19, 22, 28, 31, 32, 33, 34
dynamic_ped	7, 11, 19, 21*, 27, 31, 32, 33

The average and standard deviation after filtering out the before-mentioned data can be found in Figure 5.7. We can see that the desired effect of eliciting different levels of comfort is achieved. The calm driving style without the pedestrian is considered much more comfortable than the dynamic style with the pedestrian. It is interesting to note that the effect of the driving style seems more significant than

the pedestrian. This would contradict recent findings by Peintner et al. 2024, who also did a simulator experiment with different driving styles and pedestrian crossings while asking about the participants' desire for control, trust in automation, and acceptance. They found that the preferred driving behavior depends more on the scenario. A significant difference between their experiment and ours is that they did not have a moving base simulator, meaning that the dynamics of the driving styles were not physically experienced by the participants.

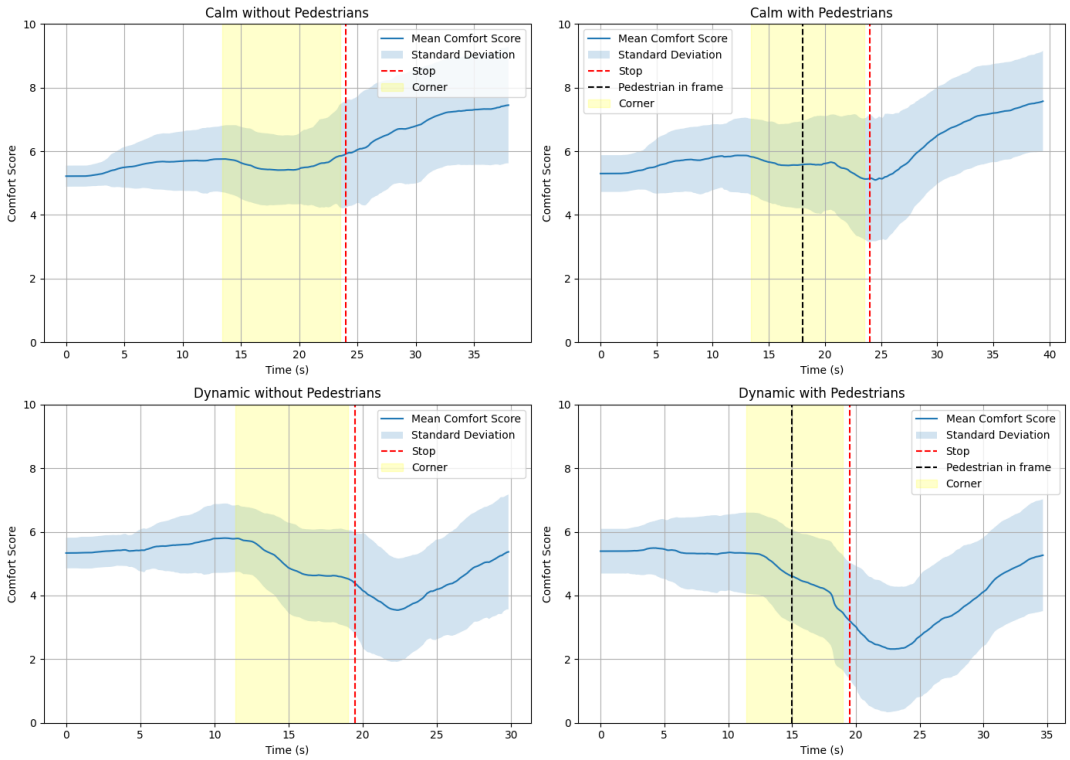


Figure 5.7: The average of the given subjective comfort over time for all runs after filtering, with the standard deviation in light blue.

A t-test was performed to see which configurations show a significant divergence from the baseline. The t-test checks if the mean of a given sample is the same as the given baseline mean. As a baseline, the average of the first 10 seconds of that run was taken, as this is the part where nothing yet happens. The results are shown in Table 5.4. All scenarios have a very significant divergence from their baseline, indicated by all p-values being below 0.01. A positive T-statistic means the mean of the given samples is greater than the baseline mean, and a negative T-statistic means the mean of the samples is smaller than the baseline mean. This shows that the calm drives cause a significant overall increase in comfort, while the dynamic drives cause a significant overall decrease in comfort. The impact of the driving style is clear.

Table 5.4: The results for the t-test of the four different configurations. The baseline is the average of the first 10 seconds. All p-values show that the divergence from the baseline in that configuration is significant. The signs of the t-statistic show that for the calm driving styles the non-baseline means are greater than that of the baseline, and for the dynamic driving styles they are smaller.

Configuration	Baseline Mean	T-Statistic	P-Value
Calm, No Pedestrian	5.30	43.9	2.05e-260
Calm, With Pedestrian	5.36	35.3	1.13e-194
Dynamic, No Pedestrian	5.38	-22.9	1.67e-93
Dynamic, With Pedestrian	5.42	-39.8	5.85e-221

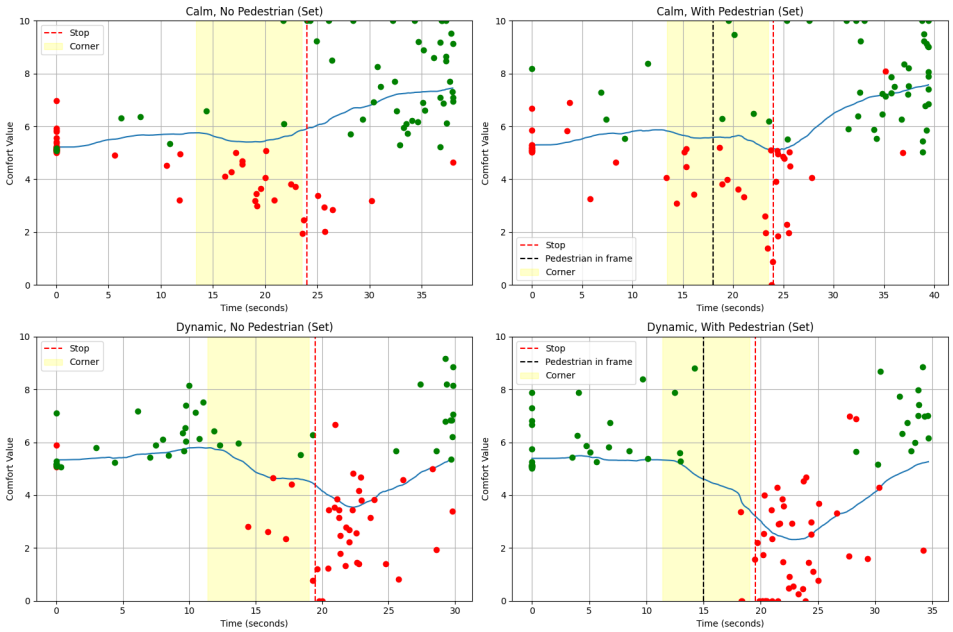


Figure 5.8: The minima and maxima are plotted as dots over the average. The red dots represent minima, and the green dots represent maxima.

Another thing to look at is the extreme values. They are plotted in Figure 5.8 where the dots represent an extreme value for one run for one participant. The red dots represent minima, and the green dots represent maxima. It is clear that there is high variety in the time of the response of the participants, as we see the dots are scattered around. Since the averaging was done per timestamp, this means that the minima of the average might not be the same as the average minima. Therefor, an analyses was done only looking at the lowest comfort level from each run. A boxplot of all the minima per configuration can be seen in Figure 5.9, to visualise the effect of the different configurations on the minima of the subjective comfort. The minima relative to their individual baselines were also calculated, these drops are presented in Figure 5.10.

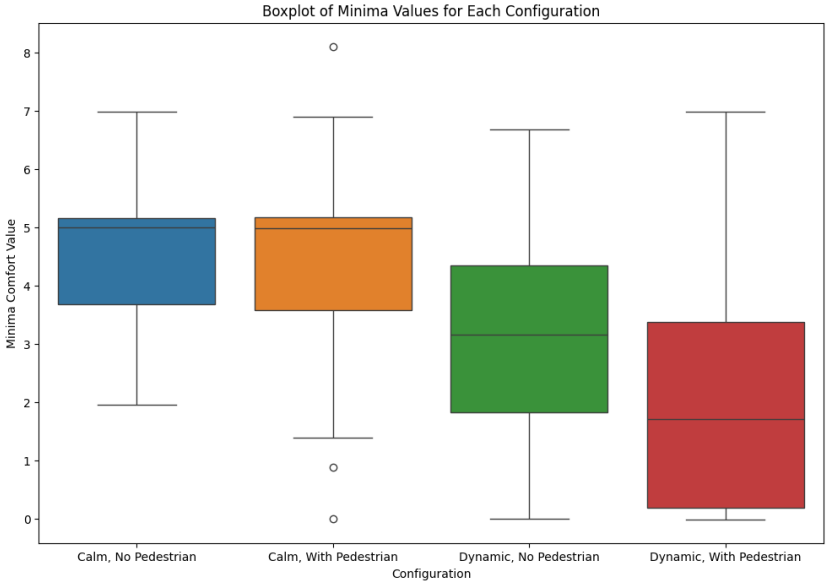


Figure 5.9: Boxplot of the absolute minima per configuration.

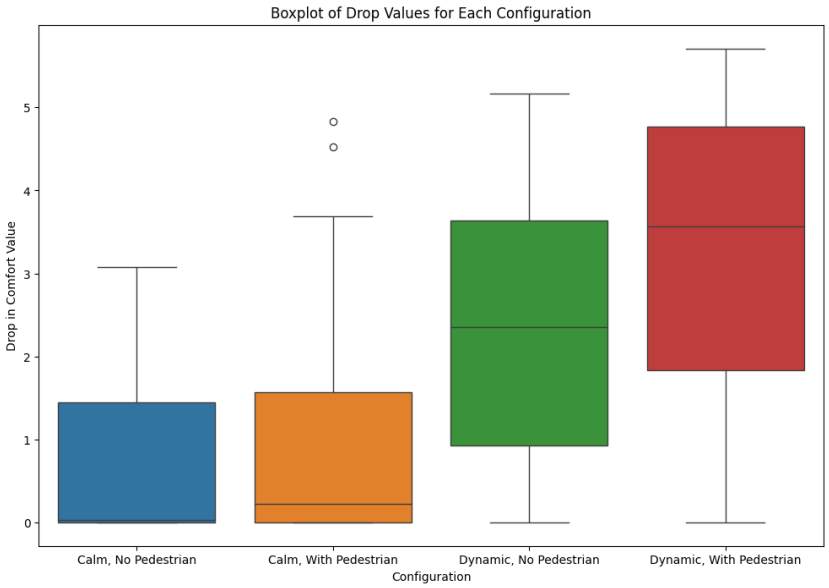


Figure 5.10: Boxplot of the drop, calculated as the minimum subtracted from the individual average of the first 10 seconds, per configuration.

To quantify the effect of the driving styles and pedestrian presence on the minimal comfort levels, both ANOVA and a pair-wise t-test test were performed on the minima, to see for which configurations they are significantly different. The f_oneway ANOVA function from the scipy library revealed an f-statistic of 26.6, and a p-value of 2.38e-14. This shows the minima of the three configurations are statistically different. The results of the pair-wise t-test are in Table 5.5. This shows that the effect of the driving style is always statistically significant on the minimal level of comfort. The effect of the pedestrian in the calm driving style is not significant though, and the effect of the pedestrian with the dynamic drive is significant ($p<0.05$) but not very ($p>0.01$). These results show that the driving style was a bigger influence than the presence of a pedestrian on the participants' minimal level of subjective comfort. Analyzing the drop values gave results that yielded the same conclusions, and those numbers are therefor left out here.

Table 5.5: Pairwise t-test between the different configurations, showing for which configurations the minima significantly differ from each other.

Comparison	T-Statistic	P-Value
Calm, No Pedestrian vs Calm, With Pedestrian	0.61	5.44e-01
Calm, No Pedestrian vs Dynamic, No Pedestrian	4.96	4.86e-06
Calm, No Pedestrian vs Dynamic, With Pedestrian	7.82	2.83e-11
Calm, With Pedestrian vs Dynamic, No Pedestrian	3.81	2.59e-04
Calm, With Pedestrian vs Dynamic, With Pedestrian	6.44	5.54e-09
Dynamic, No Pedestrian vs Dynamic, With Pedestrian	2.72	7.90e-03

It was also noted by some participants that the car braking in time for the pedestrian, as it does in the calm driving style, actually increased their trust. This is also visible in the acquired data, as we see that, even though there is a small dip in the comfort, generally the comfort goes up after the vehicle has successfully stopped for the pedestrian. This confirms what was stated by Park et al. 2022, that trust can be brought down by undesired behavior of the vehicle, but also be increased again if the vehicle shows trustworthy behavior, like braking in time for pedestrians. This emphasizes the use of making vehicles aware of and responsive to the state of the passenger.

5.1.2. Facial emotion recognition

For facial emotion recognition, the focus is on the critical scenario, which is the dynamic drive with the pedestrian in the scene. This scenario shows a clear dip in subjective comfort, so if there is an effect on the facial expression this is the scenario where it should become evident. The emotion detection model gives a confidence score to each emotion, for each of the frames in the video. The detection algorithm works frame by frame, and the webcam recorded 30 frames per second.

Figure 5.11: The moment where the pedestrian is in the visual frame for the pedestrian, this is while the vehicle is still cornering.

The results from the emotion recognition model, which was further elaborated on in Section 3.4, are plotted together with the time-to-collision, which is marked by a dashed line. The time to collision is in seconds divided by a factor of 10, to make it the same scale as the confidence scores. The moment during the right-hand turn where the pedestrian comes in the visual frame, which can be seen in Figure 5.11, is marked by a red dot. The plots can be found in Appendix D, with some shown here in Figure 5.12.

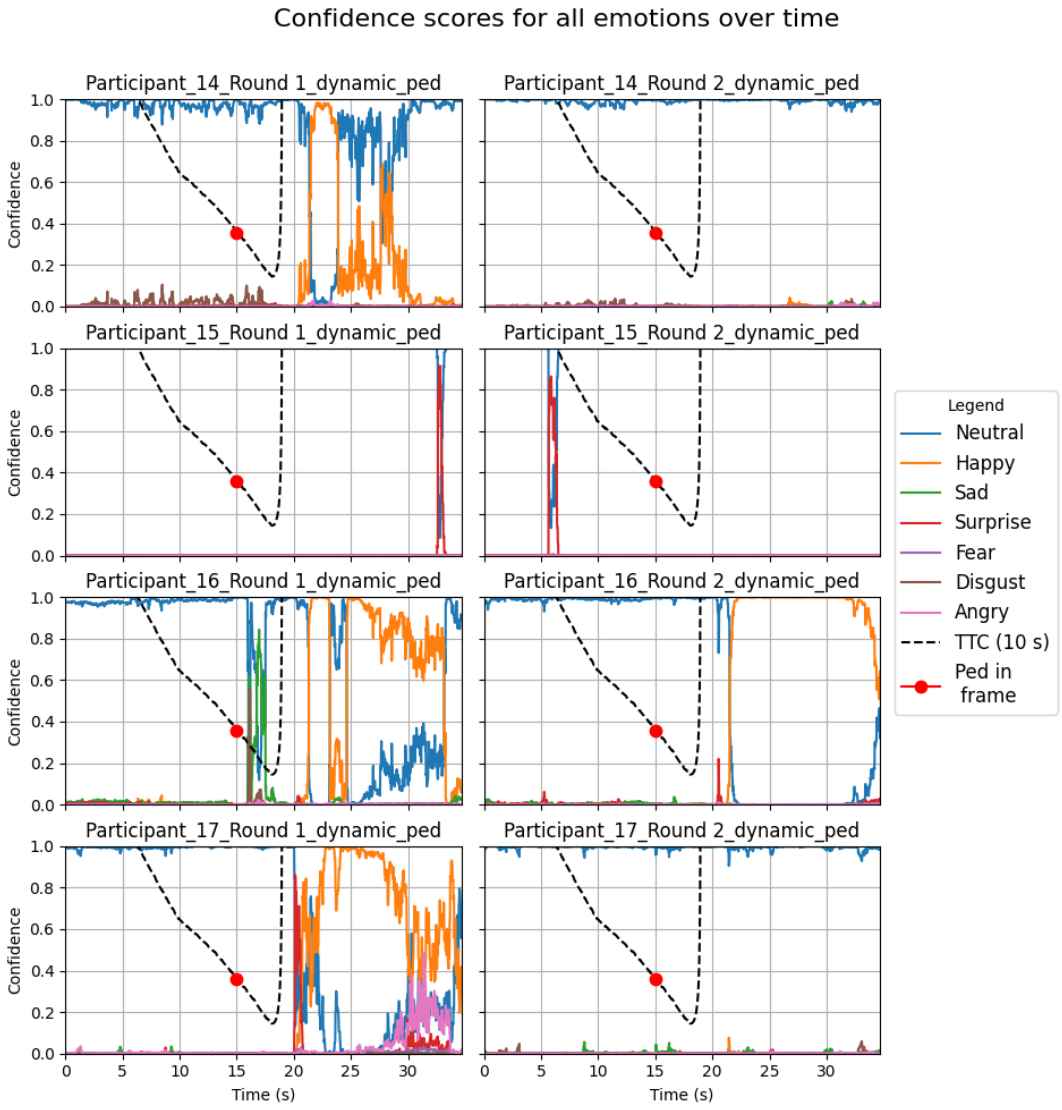


Figure 5.12: All emotions confidence scores during the dynamic scenario with the pedestrian in the scene, after applying the moving average filter, for participants 14 through 17.

Overall, Neutral is by far the most dominant emotion. This was expected. However, it was also expected to see Fear or Surprise rising at the critical event. This is not something that is consistent with the results. Two pie charts, Figure 5.13, were created to show the total percentage of each emotion, after accumulating all the confidence scores. In this pie chart we see that after Neutral, the most dominant emotions are Happy and Sad. Disgust, Surprise, Anger, and Fear all contribute very little, Fear even being the least detected emotion of all. We know from the performance of the model that it can sometimes perceive Fear as Sad, but even when we take this into account the contribution of the facial expression for Fear seems to be minimal.

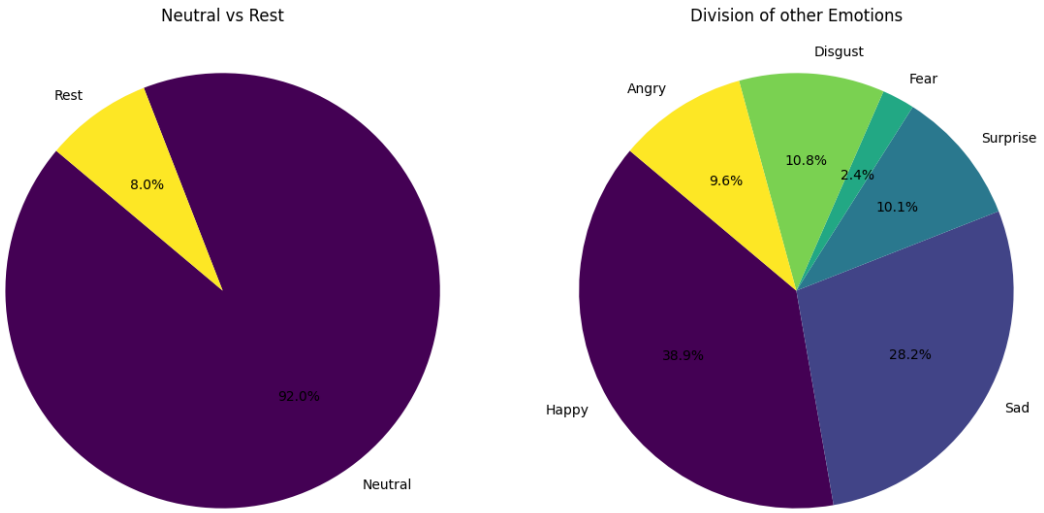


Figure 5.13: Pie charts of all accumulated confidence scores, of which the sum is 1 at every time-step. Left shows the Neutral expression versus all the others combined. The right chart shows the division of the non-Neutral labels. The percentages are (rounded to two decimals): Neutral (92.02%), Happy (3.10%), Sad (2.25%), Disgust (0.86%), Surprise (0,80%), Angry (0.77%) and Fear (0.19%).

For participant 11 the webcam footage was lost, and participant 19 was excluded for the reason that was explained in the previous section. As this is a relatively novel approach, there was not a defined way to analyze this data. The videos are not labeled, so we cannot truly check the performance of the model. However, we know its performance on the RAF-DB dataset, and the videos were analyzed together with these plots to check for coherence.

A qualitative analysis was done for each participant, looking at the results, videos, and individual frames of peaks. This analysis was described in Appendix F (not available in public version). In this appendix, the left column contains some demographic information on the participant, and whether they clearly indicate discomfort with the knob. In the middle column, a description is given of what is visible from watching the videos and the individual frames, referring also to the labels given by the emotion recognition model. Also the subjective comfort plots are presented here, with the comfort from 0 to 10 on the y-axis, and the time on the x-axis, copied from Figure 5.5. In the right column, some key frames are shown from peaks of certain emotions, or if none existed, from the neutral expression before, during, and after the critical event. These frames are also labeled with the emotion that the model predicted. It was attempted to find the frames that corresponded to the moment of the lowest comfort, so these frames also have a timestamp when relevant.

A breakdown of the different reactions detected for the participants is given in Figure 5.14. From the 32 participants:

- data was lost for one (31 left)
- one was excluded beforehand (30 left)
- two knew about the experiment, for them the pedestrian was not a surprise (28 left)
- two people were wearing double glasses which could have interfered (26 left)
- two people were less visible due to lighting conditions (24 left)

- 15 participants seem to have no significant reaction in the facial expression. (15/24)
- 8 of the participants show a Happy reaction, smiling or laughing when the vehicle has stopped in time. 4 of them also show more the expected shock reaction, which the model reads as Surprise. (8/24)
- One person showed Angry expression to the event. This seems to be an outlier. (1/24)

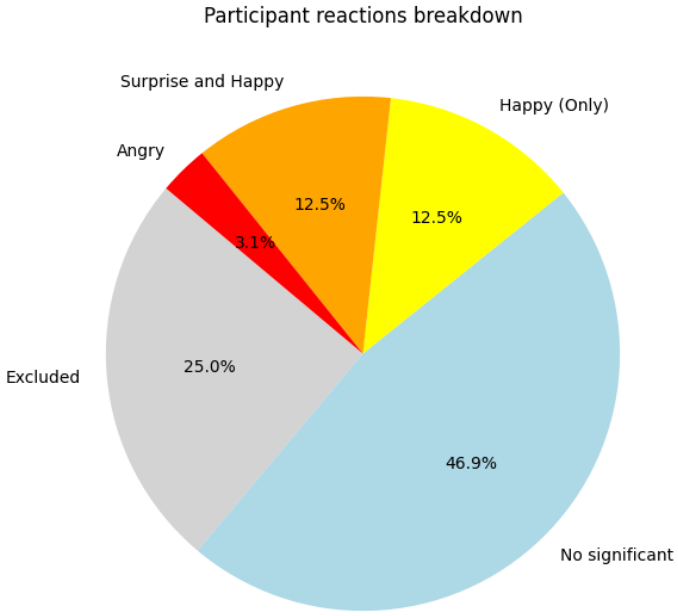


Figure 5.14: The division of the reactions detected in the participants. The total number of participants amounts to 32.

Of the 15 people that had no reaction 8 of them experienced the dynamic driving style with pedestrian first, 7 of them had already seen it in the calm driving style. From the 4 people that showed surprise followed by happy this division was 2/2, and the other 4 that showed happy also were equally divided. This suggests that the order in which the participants experienced the driving styles with pedestrian for the first time, had no influence on the facial expressions shown during the dynamic (critical) scenario.

Of the people showing a reaction, only one participant repeated this reaction with the same intensity during the second round. Of the others, half showed a similar reaction but less intense, and half did not show a similar reaction at all. This suggests that the element of not knowing what will happen is of significant influence. This is important to take into account with future research, as it is common to do repetitions in these type of experiments.

The male/female division of the dataset is 26/6. The male/female division of the 24 good measurements is 19/5. However, the division of the 15 people that showed no reaction is 14/1. One female showed the Angry expression, and of the 4 people that showed surprise and then happiness, 3 were female. Due to the low amount of female participants we can not rule out coincidence, but it could be that females are more likely to show facial expressions.

Of the participants that showed surprise + happy, 3/4 had not experienced a driving simulator. The 4 participants that showed happy all had experienced a driving simulator. Of the 15 that showed no reaction, 7 had no experience with a driving simulator and 8 did. It seems that experience with a driving simulator therefor does not matter.

The participants were also asked about their self-asserted familiarity with automated systems. The question was: How familiar are you with automated driving systems? The results of this question from

the 24 participants are in Table 5.6. The division indicates the experience is not a relevant factor in who does and does not show a reaction in the facial expressions.

Table 5.6: Division of self-asserted experience with automated driving systems over the people with and without a significant reaction in the facial expressions. The question was: how familiar are you with automated driving systems?

Experience Level	No Reaction	Reaction
Not at all	3	0
A little (read about it/seen it a few times, no personal experience)	6	4
Moderate (read about it/seen it, at least one personal experience)	2	2
A lot (studied it, multiple personal experiences)	3	3
Expert (work or have worked in the field of automated driving)	1	0

Overall we see that a majority of the participants do not have a significant response in their facial expression during this experiment, even when their subjective comfort does significantly go down. The most common reaction is that people smile or even laugh right after the most critical part (the car brakes abruptly just in time for the pedestrian) of the scenario has happened. Some of these people also show surprise just before this moment. The full individual analyses are in Appendix F (not available in public version).

5.1.3. Galvanic Skin Response

The Facial Expression analyses did not provide the desired results, even though the subjective ratings showed significant drops in comfort, in particular during the dynamic driving style. To validate if this effect was only present in the subjective ratings given by participants, or also present in the physiological data, the GSR was analysed. GSR was chosen from the available data because it has been proven to show strong correlations with arousal and stress (Dawson et al. 2007; Jaiswal et al. 2023). This effect has also been validated in the automotive context (Wang, Murphey, et al. 2019; Memar et al. 2021).

Before analysis, the data was inspected. From one participant, the data showed a response that was unlikely to be physiological. This response, of which one run is shown in Figure 5.15, showed abnormally fast and large drops in the GSR, also showing response amplitudes much higher than 3 µSiemens (Braithwaite et al. 2013). Something else that was noticed, was that a lot of responses were relatively low. Normal GSR values do not go below 1µSiemens (Braithwaite et al. 2013), which did occur in this dataset. A problem could have been that the velcro straps were not tight enough, resulting in the electrode to momentarily lose contact with the skin. A cut-off was implemented, excluding participants whose data dropped below 1 µSiemens, even if it was momentarily because the full response was deemed no longer reliable. This resulted in excluding 6 participant, which is a lot considering it is close to 20% of the data. To avoid this in the future, the one-time-use sticky electrodes could be used.

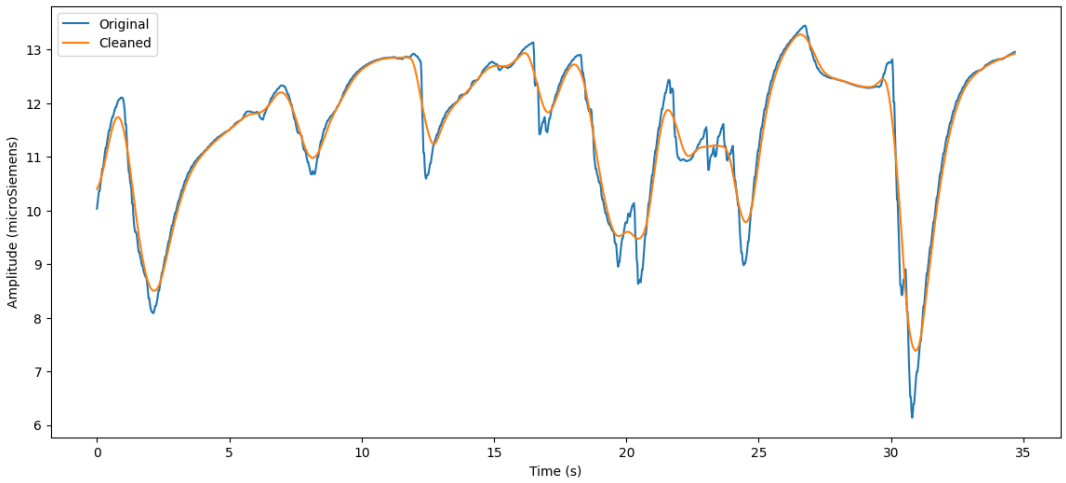


Figure 5.15: Faulty GSR with extremely big drops in a short amount of time.

The most significant variable of the experiment was found to be the driving style, and to not let the validation of GSR depend on the individual subjective comfort scores, an event-based analysis was done. As explained in Chapter 3.3, the phasic component, or the Skin Conductance Response, is known to be an indicator of events that cause arousal or stress. Peak detection was performed, resulting in dictionaries with onsets, the moment in time that the peak starts, peaks, the moment the phasic component is at its highest, and the amplitudes, the difference between the peak value and the onset value. This detection is also depicted in Figure 3.5.

As features, the times and amplitude of the highest peak for each run were given to a Random Forest Classifier. The classifier then predicted if this was a calm or dynamic drive. Because results for such classifiers can depend on the random seed that is used, the fitting and prediction were repeated 10 times with a different seed. The classifier could always reach a perfect fit on the training data, with a mean accuracy of 100% and a standard deviation of 0. On the testing data, the mean accuracy was 74.3% with a standard deviation of 4.7%. This analysis was not fine-tuned, as the goal was merely to show that GSR can be utilized to classify comfortable or uncomfortable events in this dataset. By fine-tuning further, both in feature selection and the classifier itself, it is likely possible to achieve better results.

These results confirm that the negative response in comfort during the dynamic driving style, though not visible in the facial expressions of most participants, is present in the physiological data.

5.2. Neural network for Comfort Prediction

Because the facial expressions did not provide the results that we hoped for, it was deemed not possible to use that data to attempt to set up a prediction model. The influence of the driving style and the presence of a pedestrian on the subjective comfort was shown, and it was shown that GSR could already be used to differ between the two driving styles, which suggests that GSR could also be used for individual comfort prediction. For this, the data was inspected again, and both the participants with faulty GSR and the runs with inconsistent subjective ratings were excluded.

As elaborated in Chapter 3.3, GSR is split up in the Tonic and Phasic components. Both the Tonic component and the Phasic component are extracted as possible features for the model to train on, together with the time-to-collision (ttc) and longitudinal and lateral accelerations. As was presented in Figure 3.9, the GSR data, ttc, and accelerations are each passed through their own feature extraction models consisting of three Convolutional layers and two LSTM layers. The outputs of these models are then combined and an attention layer is applied. Finally, a fully connected layer is used to get a prediction for the comfort value. The mean absolute error of the prediction versus the true value is used as the loss function to train the model. The model was trained to give one output for a sequence corresponding to 5 seconds of data in 10Hz, where the true label was the comfort score at the end of those 5 seconds. These comfort scores were available in 2 decimal precision but rounded to 1 decimal, as turning the knob by a smaller amount on purpose was practically impossible. It was chosen not to round the labels to integers, because changes from for example 4.45 to 3.50 would have been lost in the data.

In pre-processing, the data was resampled to 10Hz and then the sequences of 5 seconds were created. The data was then standardized using a standard scalar, which gives the data zero mean and a standard deviation of 1. For the GSR data, this was not done on all data combined, because this would make the GSR signal for people with a low mean almost zero. For this reason the GSR signals were standardized individually in the signal processing, before splitting it into the tonic and phasic components. 20% of the data is set aside as a validation set, and 80% is used as training data. In this split it is made sure that both sets contain roughly the same distribution for the labels. This can be seen in Figure 5.16. We also see that the dataset is not balanced, with much less data near the extremes. To account for this, different strides were applied when sampling the data. The stride of the sequence sampling is how much the window shifts to get the next sequence. This stride was made dependent on the label. For the labels that were far from the center a smaller stride was applied, and for the more common labels a larger stride was applied. This balanced the data out more. As minimum stride 0.5 seconds was chosen, which is 10% of the sequence length, to avoid different sequences from being too much alike which would result in overfitting. This stride of 0.5 seconds was applied to sequences with a label below 3 and above 8, as these are the most extreme regions with the lowest amount of samples. For labels between 3 and 5, and between 6 and 8, a stride of 2 seconds was

applied. For the very common region between 5 and 6, a stride of 4 seconds was applied. These different strides resulted in a division as can be seen in Figure 5.17. It is clear that the labels are now represented much more equal in the dataset.

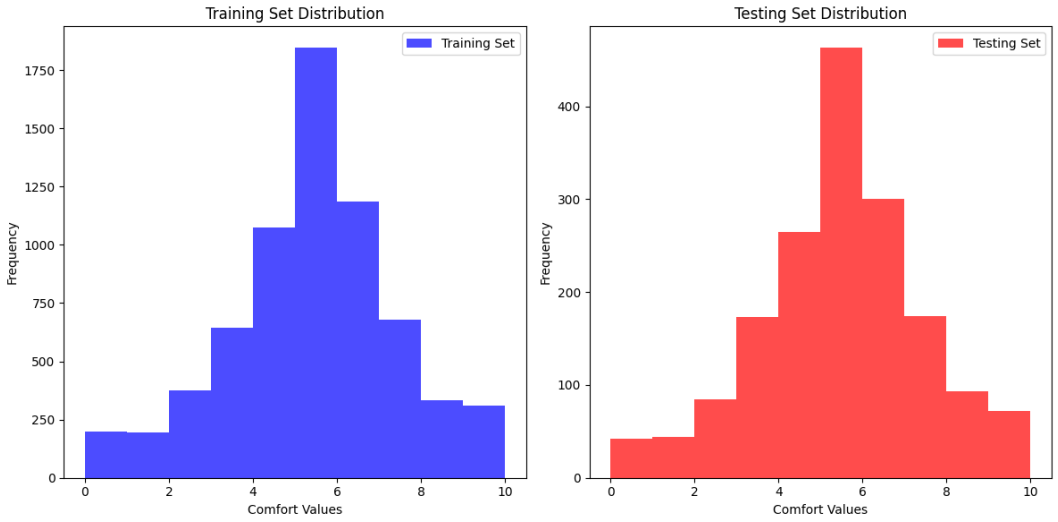


Figure 5.16: Data distribution of the training and validation set, when sampling with a constant stride of 0.5 seconds.

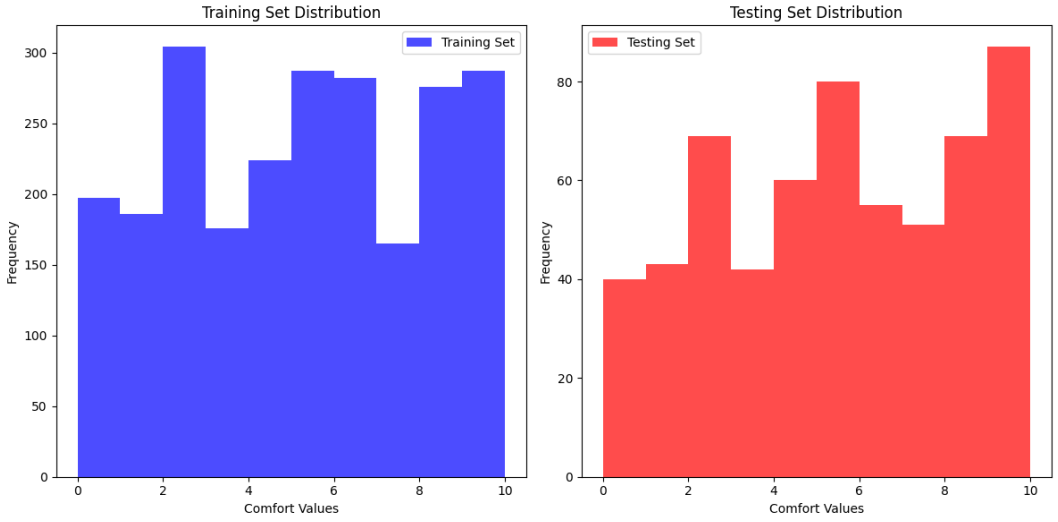


Figure 5.17: Data distribution of the training and validation set, after applying different strides to different regions. The distribution is much more equal.

During training, after each epoch, the performance of the model on the testing set was calculated. If that performance was better than what was previously registered as the best performance, the weights from that epoch were saved. At the end of the 500 epochs, the best model weights are loaded to acquire the results. During training the intermediate results are also stored, plotting these shows the training progress of the model, confirming that it is actually learning and improving over time. In Figure 5.18, the progress of the model trained on vehicle dynamics and GSR data for 500 epochs can be found.

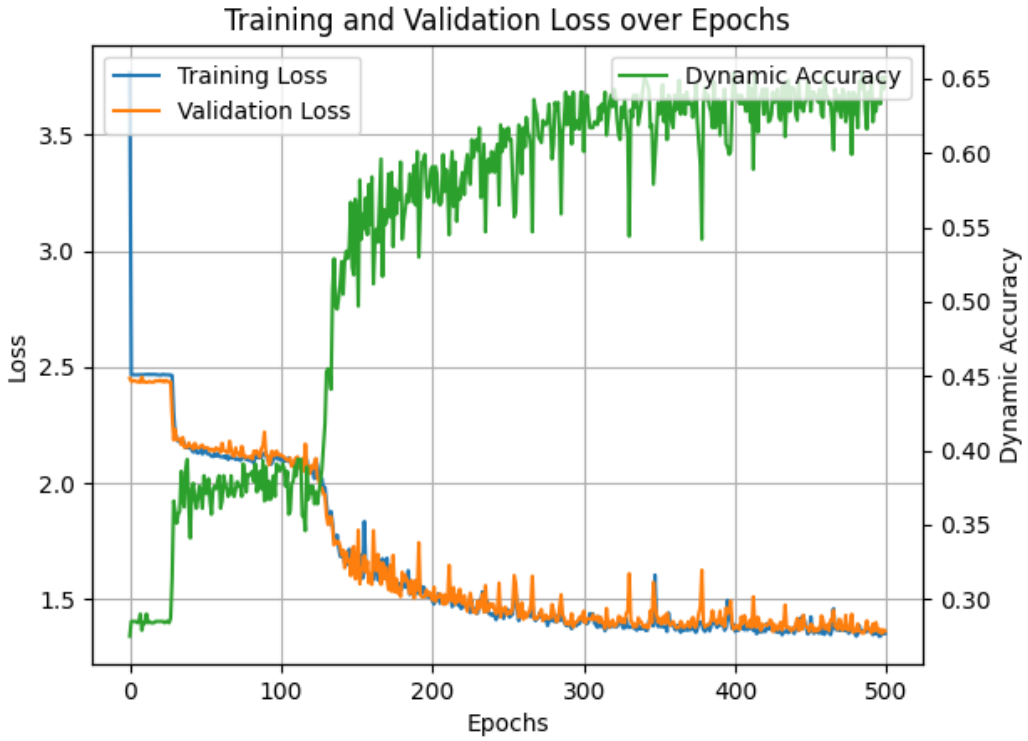


Figure 5.18: Loss of the model trained on vehicle dynamics and GSR components for 500 epochs.

The performance of the model was compared for the different features, looking at the loss (mean absolute error) and accuracy on both the fitted training set and the validation set. For the accuracy, a margin of error was given, since it can be considered correct if the model predicts a value close to the label. Also, the true labels in this case are the subjective scores from the participant, which can be subject to biases discussed in Chapter 2.3. This margin of error was chosen to be plus or minus 1 around label 5, and linearly building up to 2 on both outer labels, 0 and 10. It was selected like this because at the edges the exact number becomes less important, a comfort rating of 0 and 2 both mean extremely uncomfortable. Near neutral it should be a bit more exact since the difference between 4 and 6 means uncomfortable versus comfortable. A plot showing this area with the results of the validation set for the model trained on can be found in Figure 5.19

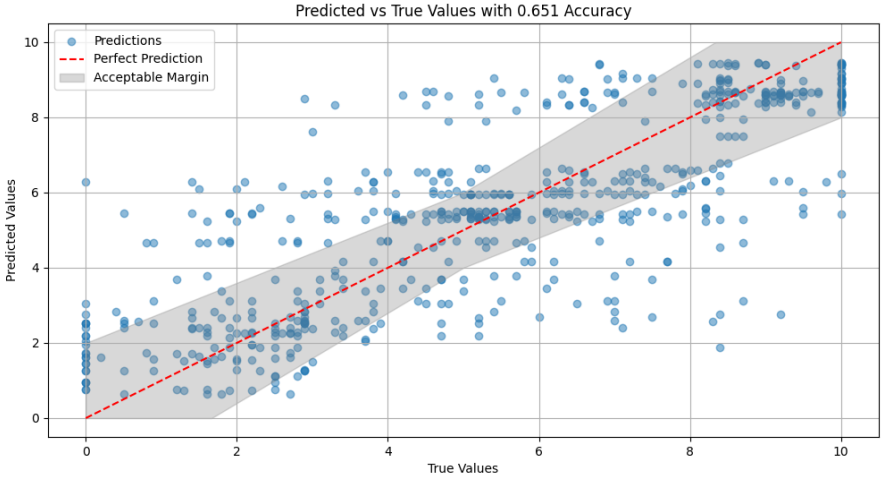


Figure 5.19: Performance on the testing set. 65.1% of the samples are within the set margin.

The different feature combinations are compared in Table 5.7. We see that in terms of accuracy, the best performance is the model that is trained on both vehicle dynamics and GSR. However, only vehicle dynamics does reach a slightly lower mean absolute error, although this difference is very small with the model that combines the two. In terms of fitting the model to the training data, using only vehicle dynamics seems to work the least well, and only GSR makes a better fit than the combined model. Here it should be taken into account, that due to this being from a simulation, the accelerations are the same for every run, meaning that also between the training and testing data, there will be a lot of highly similar samples, if not exactly the same. This is generally not desired, cause it will make the model less applicable to data outside of this experiment, thus it is not possible to really test its validity. Another thing that can be noted is that because the labels are from different people, giving different ratings, all while experiencing the same accelerations, the model gets exactly the same input, but with different output labels to converge to. This limits the extend to which the model can fit to the data. This was confirmed when trying to run the model for 1000 epochs, which improved the fitting on the vehicle dynamics data only slightly both for the MAE (-0,023) and the accuracy (+1.4%). When doing the same with the model only trained on GSR data, a much bigger improvement was found on both the MAE (-0,248) and accuracy (+7.1%). This shows that to predict individual comfort levels, you want to incorporate physiological data, and GSR is a viable predictor. TTC is another feature that is always the same, and also not present in two of the runs. It can be seen that this has no positive effect on the performance, it just results in more overfitting. TTC was thus left out in further processing.

GSR could be used to discriminate between the two driving styles, which was shown to be the biggest factor between comfortable and uncomfortable drives, and it shows promising results for predicting individual comfort levels.

Table 5.7: Performance for different features. VD stands for vehicle dynamics, which include longitudinal and lateral accelerations. GSR consists of two features, the tonic component and the phasic component. TTC is time to collision.

Features	train MAE	test MAE	train Accuracy	test Accuracy	Avg Correlation
VD	1,306	1,330	64,8%	62,9%	0.61
GSR	1,234	1,660	70,6%	55,5%	0.21
VD + GSR	1,321	1,352	65,4%	65,1%	0.61
VD + GSR + TTC	1,029	1,426	76,2%	60,9%	-

To better validate the performance of the model on individual comfort prediction, it had to be tested on unseen data. Because due to the overlap in the sequences (stride < window) some sequences in the validation set might have been similar to sequences from the training data. Also, we want to see the performance on individual data and not just the full dataset. To achieve this, Leave-One-Out-Cross-Validation (LOOCV) was applied. First, the data from one participant was completely excluded from the

dataset. The model was then trained exactly as before, but now an extra validation was applied. The data of this excluded participant was given in original order, meaning we could see how the prediction of the model would follow the trend of the original subjective comfort scores, as presented in Figure 5.20.

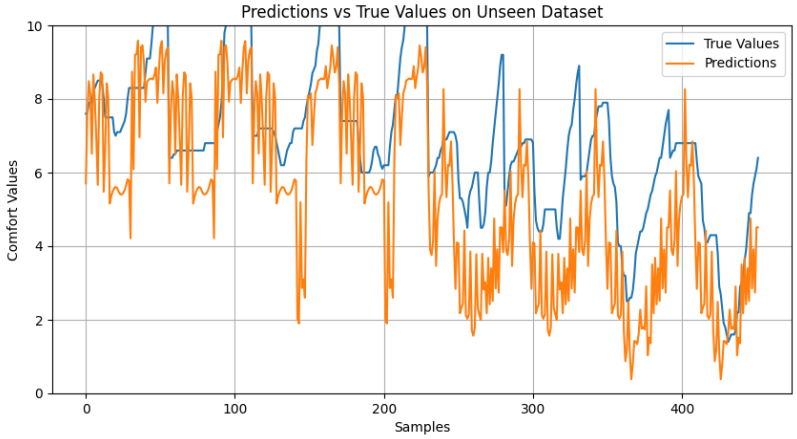


Figure 5.20: True values and predictions on the unseen data of participant 23, from the model trained on both vehicle dynamics and GSR data.

The results of this LOOCV were analyzed by looking at the correlation between the predicted values and the true labels of the unseen participants' data. The averages are also added to Table 5.7. Here the combined model clearly performs better than the individual features. For the GSR only model, all correlations were positive, and 11/14 were significant ($p<0.01$). For only vehicle dynamics, all correlations were significant and much stronger than for the gsr. For the combined feature model, all 14 correlations were positive and significant, showing highly similar performance to the model with only vehicle dynamics. This is an indication that the model still fits itself on the dynamics data since this is always the same input, and the effect of the GSR is therefore minimal. All plots of the combined model can be found in Appendix H.

6 Conclusion

The upcoming developments in automated driving promise to increase the quality of our daily lives with fewer traffic accidents, fewer conjunctions, and more time that can be spent on other activities. However, these benefits rely on large acceptance by the public, which research has shown to be very low to be very low at this moment (Su et al. 2023; Park et al. 2022; Mara et al. 2022; Nordhoff et al. 2023). This is why understanding comfort in automated vehicles is critical for further development in this area. This research contributes to the ambition of acquiring objective methods for assessing the emotional state of the drivers or passengers in automated vehicles.

In this thesis, an experiment was set up and carried out on a driving simulator to let people experience a drive by an automated vehicle to elicit different levels of comfort by varying the driving style and the presence of a pedestrian while collecting a comprehensive dataset to analyze these levels of comfort. This experiment was successful, and a summary of the contributions is given in Section 6.1. The research question and hypotheses that were introduced in Chapter 1 are revisited in Section 6.1.1. As this work does not mark the end of research in the area of comfort in automated vehicles, recommendations for future research are given in Section 6.2. Finally, a personal vision for the future of this research topic is given in Section 6.3.

6.1. Summary of the Contributions

A comprehensive dataset was collected from a vehicle simulator experiment that elicited different levels of comfort in terms of trust and perceived risk. This dataset includes continuous subjective comfort ratings given by the participant, vehicle dynamics from the real-world drive on which the simulation was based, webcam footage monitoring the person's facial expression, Galvanic Skin Response, heart rate (variability), and eye-tracking from 32 participants. Such comprehensive datasets are rare in literature and allow for more future research in comparing the different signals and their relation. Setting up and executing an experiment like this is highly time-consuming, and human data is generally difficult to acquire. This dataset is therefore a substantial contribution to future research, and the experience gained from this experiment also provides a basis for recommendations for future work.

The subjective comfort ratings were analyzed and their relation to the different driving styles and the presence of a pedestrian were quantified. With this also the viability of the experiment is shown, as it shows that different levels of comfort are elicited by using different driving styles and adding or removing other road users. Also, it was shown that in this setup the driving style was more important for the comfort than the presence of a pedestrian.

After investigating different publications, a state-of-the-art model by Zhang et al. 2023 was successfully implemented for this purpose. With minimal lighting conditions, where the only source of light was the monitor in the simulator, the face could always be detected, and expressions were successfully detected and given corresponding emotion labels. This shows it is possible to implement such models in the context of automated driving.

The results from the facial expression recognition model were analyzed to assess its viability as a method for comfort assessment. This model labeled the data frame by frame with an emotion from the widely accepted set of basic emotions: Neutral, Happy, Sad, Angry, Disgust, Surprise, and Fear. Each

video and the model output was then individually analyzed, to see what was in response to the event. Even though this did not provide the expected and aspired results, which will be further elaborated in Section 6.1.1, it does provide a basis for future research.

A neural network was implemented to predict the subjective comfort of a person from vehicle dynamics and their GSR. It was shown that both GSR and vehicle dynamics can successfully fit on the dataset, where vehicle dynamics shows better performance on the test set, but also limitations on how well it can fit the training set. More importantly, the LOOCV analyses showed the model was able to follow trends of subjective comfort well for data from an unseen participant, and for all participants, the self-asserted comfort and the predicted comfort show a positive correlation. This was the case both with and without including the GSR Features. This might be caused by the fact that the vehicle dynamics are highly correlated with the subjective comfort in this experiment, and since the dynamics are always the same the model learns this relation and the varying GSR signal becomes of little influence. This model was taken from literature and not further fine-tuned, meaning it is likely possible to get better results from this dataset in future research. These findings do already show the potential of achieving objective comfort assessment in automated vehicles, losing the biases that are inherent to subjective assessment, and paving the way for future studies in this area.

6.1.1. Research question & hypotheses

The main research question was: How can facial emotion recognition benefit the assessment of comfort, concerning trust and perceived safety, in automated vehicles? After executing the experiment and analyzing the data, the data, and the corresponding results from this analysis, were deemed to be inadequate to provide a confident answer to this question. The formulated hypotheses were:

- H1: Including facial emotion recognition in the dataset will improve the performance in terms of correctly predicting the individual comfort level, compared to only using the vehicle data. Because comfort is subjective, it differs per person per situation, and the vehicle data does not contain this information.
  - H1.1: The biggest difference in performance will be seen in extreme cases, the cases that diverge the most from the average. With this, it is meant that for participants that are overall much more or less comfortable than the average, the performance will increase the most by including facial emotion recognition.
- H2: Events that cause most subjective discomfort, will show the biggest response in facial expressions.
- H3: Fear and surprise will have the strongest, negative, correlation with feeling comfortable. When the vehicle behaves not as expected or in an unsafe way, these are the emotions expected to be seen, and the comfort is expected to decrease.

Going from this dataset, H1 and H2 have to be rejected. Most participants (15 out of 24) did not exhibit any changes in their facial expressions during the most critical event, even when their subjective comfort levels decreased significantly, and a response in GSR was also found. This lack of response indicates no support for the notion that facial emotion recognition improves performance in detecting discomfort. H1.1 is also rejected after the analyses done in Section 5.1.2, as there was no clear distinction between participants that did or did not show responses in the facial expression related to their subjective comfort scores.

H3 is partially rejected, as Fear was actually the least present emotion throughout the full experiment (see Figure 5.13). However, as was depicted in Figure 5.14, amongst the participants that showed a response in their facial expression as a reaction to the critical event, surprise was detected for half of them, which is in favor of this hypothesis.

The expected emotions associated with a decrease in comfort, such as surprise and fear, were rarely captured by the facial expression recognition model during the experiment, with only 4 participants showing such an expression at the moment of the critical event. Although a visual reaction occurred for 9 of the participants, for 8 of them this included smiling, which resulted in ”Happy” being registered as the dominant emotion. This response is counter-intuitive, as happiness is typically linked to an increase in comfort rather than a decrease. However, the presence of this reaction, combined with the fact that it does not consistently recur once the pedestrian is no longer a surprise, suggests that further

investigation in a real-world setting could be valuable. It might be the case that the participant would need to momentarily feel as if there is a genuinely dangerous situation. This makes it difficult to test.

The unexpected elicitation of happiness is likely due to the scene's lack of realism and the participants' awareness that they were in a simulator, which diminished any sense of real danger. The difference between the two rounds for many of the participants who showed this reaction indicates that the element of surprise was relevant to their response. Additionally, camera placement and lighting conditions may have influenced the results. The webcam was located above the screen, which for most participants meant it was quite a bit above eye height. This plus the fact that participants would have their head tilted forward a bit and were wearing large eye tracking glasses, meant the eyes especially were not always very well visible. Although the face was always detectable, obscuring parts of the face, especially in the regions where a lot of features are extracted like around the eyes, decreases the performance of the emotion model as also noted by Zhang et al. 2023. Furthermore, the monitor was the sole light source, sometimes causing reflections in the glasses depending on the participants' head pose, further complicating the emotion detection process.

Overall, the conclusion is that FER is not successful as a method for comfort assessment in automated vehicles. We are also aware of other ongoing studies at this time that have similar findings, suggesting that it is not solely inherent to this specific experimental setup.

6.2. Future work & Recommendations

Current state-of-the-art models can successfully classify emotions by analyzing facial expressions. However, a lot of research in the area is based on laboratory settings, or labeling pictures from the internet (Li, Deng 2020). These pictures are often posed, or spontaneous expressive pictures are chosen. It seems that this does not make them applicable in daily life scenarios, as most of the time the so-called display rules are in effect. These "rules" are generally learned by someone, and can depend on sex, culture, age, country of origin, and other demographics (Ekman, Friesen 1969). They can determine how much we show certain emotions on our faces in certain situations. In this experiment, we saw a lot of people not showing any emotional expression, even when their subjective ratings would say they felt uncomfortable. Going forward, it would seem that other measures like GSR are more reliable as a reference to someone's state. Nevertheless, there are some possibilities to further explore facial expressions.

In future research, it could be interesting to see an analysis of facial expressions in a real-world drive, or a more realistic simulation environment. Using a moving base simulator has proven to be very important when testing different driving styles, as the driving style was found to have more influence than the presence of a pedestrian during this experiment. This contradicts other recent research that found road users to be more influential than driving style (Peintner et al. 2024). They had a much more realistic environment using VR, but not a moving base simulator. Combining these could bring out even more accurate results. However, using VR would likely make it impossible to look at facial expressions. So either a realistic setup must be created without using VR, or a real-world test should be done, to test the potential of facial expression recognition in more realistic conditions.

In this research, it was chosen to use the discrete classification method, because more datasets and pre-trained models were available for this. However, looking at the other emotion detection methods is something that future research should look into. For example, another popular approach instead of the discrete classification is defining the emotional response in terms of arousal and valence, meaning the intensity and pleasantness of the experienced emotion respectively. If a facial expression dataset with these labels in the context of automated driving were to be developed, this would help continue this research a lot. Emotional state in terms of valence and arousal in automated vehicles is already investigated in the literature using other methods, like eye tracking (Mou, Zhao, et al. 2023). However, it should be noted that the display rules might again come into play, with people (subconsciously) limiting the arousal or valence that they show in their expression.

Another approach that has shown promise in using facial expression recognition for passenger discomfort detection was done by Beggiato, Rauh, et al. 2020. Instead of classifying emotions, they just detect different Action Units, form the Facial Action Coding System, and correlate those in an uncomfortable scenario in a driving simulator. They find correlations for certain Action Units, like raising eyebrows and the upper lip, that are generally also linked to a surprised expression Beggiato, Rauh, et al. 2020. Their analyses were done completely on the aggregated data, so future research could

look into whether this is also viable on an individual level.

A combination of GSR and facial expressions was also investigated by Meza-García et al. 2021. GSR is used to assess arousal, however, high arousal can be both positive and negative. They argue that facial expressions can be used to determine which it is. In this thesis, it became evident that during simulator experiments it can occur that people smile while still feeling negative comfort. This means that even when GSR indicates high arousal at that moment, the facial expression model would mark this as positive, negatively impacting the state estimation of the participant. Further research could be done to confirm this.

In future experiments where subjective ratings are still taken, the researchers should look into developing a device that is less of a distraction from the experiment. The turning knob can be cause for people to look down and away from the scene, even when explicitly instructed to pay as little attention to it as possible. Developing a device where it is not necessary to look at the device itself would make it much less distracting. Overall, finding objective methods for the comfort assessment of participants is the most desirable. This not only avoids a distraction during the experiment but also avoids the many biases that are inherent to the subjective methods. At the moment of writing the people at Siemens are already doing this by integrating a spring in their knob, which gives the person direct haptic feedback, for future experiments.

6.3. Final Thoughts

It is essential to consider that the vehicles of the future have to combine multiple inputs. If certain changes in emotions are registered when the car also registers situations that are potentially critical, like pedestrians crossing or sudden accelerations, it is likely that this emotional response is a reaction to what is happening. Also, other physiological data can be utilized. It would be interesting to see if this can already be done with current everyday wearable devices, like smartwatches. Since it is unlikely that people will attach all the electrodes, as is done in this experiment and many others, every time they want to go for a drive. But many people already own smartwatches, and possibly it could even be an item that comes with the vehicle. Going a step further it might even function as the the car-key. All in all, the transition to automated driving is moving ahead at a high pace, and it is critical that we take the comfort of the intended users in high regard with these developments. Automated driving shows great potential, and while technological advancements are making great steps, the objective now is to increase their acceptance by the general public.

References

AbuAli, Najah, Hatem Abou-Zeid (2016). “Driver behavior modeling: Developments and future directions”. In: *International journal of vehicular technology* 2016.

Aggarwal, Charu C et al. (2018). “Neural networks and deep learning”. In: *Springer* 10.978, p. 3.

Beggiato, Matthias, Franziska Hartwich, et al. (2020). “KomfoPilot—Comfortable automated driving”. In: *Smart automotive mobility: reliable technology for the mobile human*, pp. 71–154.

Beggiato, Matthias, Nadine Rauh, Josef Krems (2020). “Facial expressions as indicator for discomfort in automated driving”. In: *Intelligent Human Systems Integration 2020: Proceedings of the 3rd International Conference on Intelligent Human Systems Integration (IHSI 2020): Integrating People and Intelligent Systems, February 19-21, 2020, Modena, Italy*. Springer, pp. 932–937.

Bellem, Hanna et al. (2018). “Comfort in automated driving: An analysis of preferences for different automated driving styles and their dependence on personality traits”. In: *Transportation research part F: traffic psychology and behaviour* 55, pp. 90–100.

Bhide, Neeraja et al. (2023). “Defining, measuring, and modeling passenger’s in-vehicle experience and acceptance of automated vehicles”. In: *arXiv preprint arXiv:2309.10596*.

Blanco-Claraco, J. L. (2021). “A tutorial on SE(3) transformation parameterizations and on-manifold optimization”. In: *arXiv preprint arXiv:2103.15980*.

Braithwaite, Jason J et al. (2013). “A guide for analysing electrodermal activity (EDA) & skin conductance responses (SCRs) for psychological experiments”. In: *Psychophysiology* 49.1, pp. 1017–1034.

Cai, Hua, Yingzi Lin (2011). “Modeling of operators’ emotion and task performance in a virtual driving environment”. In: *International Journal of Human-Computer Studies* 69.9, pp. 571–586.

Calvo, Manuel G, Daniel Lundqvist (2008). “Facial expressions of emotion (KDEF): Identification under different display-duration conditions”. In: *Behavior research methods* 40.1, pp. 109–115.

Carreiras, Carlos et al. (2022). *BioSPPy: Biosignal Processing in Python*. [Online; accessed 25/07/24]. URL: https://github.com/PIA-Group/BioSPPy/.

Castellanos, Juan C, Fabiano Fruett (2014). “Embedded system to evaluate the passenger comfort in public transportation based on dynamical vehicle behavior with user’s feedback”. In: *Measurement* 47, pp. 442–451.

Chen, Xiaobo et al. (2023). “Transferable driver facial expression recognition based on joint discriminative correlation alignment network with enhanced feature attention”. In: *IET Intelligent Transport Systems*.

Cieslak, Maciej et al. (2020). “Accurate ride comfort estimation combining accelerometer measurements, anthropometric data and neural networks”. In: *Neural Computing and Applications* 32, pp. 8747–8762.

Cohn, Jeffrey F, Zara Ambadar, Paul Ekman (2007). “Observer-based measurement of facial expression with the Facial Action Coding System”. In: *The handbook of emotion elicitation and assessment* 1.3, pp. 203–221.

Committee, Vehicle Dynamics Standards (2022). “Vehicle Dynamics Terminology J670_202206”. In: *SAE International*, p. 73.

Davoli, Luca et al. (2020). “On driver behavior recognition for increased safety: a roadmap”. In: *Safety* 6.4, p. 55.

Dawson, Michael E, Anne M Schell, Diane L Filion, et al. (2007). “The electrodermal system”. In: *Handbook of psychophysiology* 2, pp. 200–223.

Deng, Jiankang et al. (2019). *RetinaFace: Single-stage Dense Face Localisation in the Wild*. arXiv: 1905.00641 [cs.CV]. URL: https://arxiv.org/abs/1905.00641.

Deubel, C, S Ernst, G Prokop (2023). “Objective evaluation methods of vehicle ride comfort—A literature review”. In: *Journal of Sound and Vibration* 548, p. 117515.

Du, Guanglong et al. (2020). “A convolution bidirectional long short-term memory neural network for driver emotion recognition”. In: *IEEE Transactions on Intelligent Transportation Systems* 22.7, pp. 4570–4578.

Egmont-Petersen, Michael, Dick de Ridder, Heinz Handels (2002). “Image processing with neural networks—a review”. In: *Pattern recognition* 35.10, pp. 2279–2301.

Ekman, Paul et al. (1999). “Basic emotions”. In: *Handbook of cognition and emotion* 98.45-60, p. 16.

Ekman, Paul (n.d.). *Universal Facial Expressions*. Accessed: 2024-06-20. URL: https://www.paulekman.com/resources/universal-facial-expressions.

Ekman, Paul, Wallace V Friesen (1969). “The repertoire of nonverbal behavior: Categories, origins, usage, and coding”. In: *semiotica* 1.1, pp. 49–98.

— (1971). “Constants across cultures in the face and emotion.” In: *Journal of personality and social psychology* 17.2, p. 124.

— (1978). “Facial action coding system”. In: *Environmental Psychology & Nonverbal Behavior*.

Ekman, Paul, E Richard Sorenson, Wallace V Friesen (1969). “Pan-cultural elements in facial displays of emotion”. In: *Science* 164.3875, pp. 86–88.

Fabian Benitez-Quiroz, C, Ramprakash Srinivasan, Aleix M Martinez (2016). “Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5562–5570.

Fritsch, F. N., J. Butland (1984). “A Method for Constructing Local Monotone Piecewise Cubic Interpolants”. In: *SIAM Journal on Scientific and Statistical Computing* 5.2, pp. 300–304.

Golding, John F (2016). “Motion sickness”. In: *Handbook of clinical neurology* 137, pp. 371–390.

Goodfellow, Ian, Yoshua Bengio, Aaron Courville (2016). *Deep Learning*. http://www.deeplearningbook.org. MIT Press.

Goodfellow, Ian J, Dumitru Erhan, et al. (2013). “Challenges in representation learning: A report on three machine learning contests”. In: *Neural information processing: 20th international conference, ICONIP 2013, daegu, korea, november 3-7, 2013. Proceedings, Part III 20*. Springer, pp. 117–124.

Greco, Alberto et al. (2015). “cvxEDA: A convex optimization approach to electrodermal activity processing”. In: *IEEE transactions on biomedical engineering* 63.4, pp. 797–804.

Haboucha, Chana J, Robert Ishaq, Yoram Shiftan (2017). “User preferences regarding autonomous vehicles”. In: *Transportation research part C: emerging technologies* 78, pp. 37–49.

Harb, Mustapha et al. (2021). “What do we (Not) know about our future with automated vehicles?” In: *Transportation research part C: emerging technologies* 123, p. 102948.

He, Xiaolin et al. (2022). “Modelling perceived risk and trust in driving automation reacting to merging and braking vehicles”. In: *Transportation research part F: traffic psychology and behaviour* 86, pp. 178–195.

Hoff, Kevin Anthony, Masooda Bashir (2015). “Trust in automation: Integrating empirical evidence on factors that influence trust”. In: *Human factors* 57.3, pp. 407–434.

Holtgraves, Thomas (2004). “Social desirability and self-reports: Testing models of socially desirable responding”. In: *Personality and Social Psychology Bulletin* 30.2, pp. 161–172.

Holthausen, Brittany E, Rachel E Stuck, Bruce N Walker (2022). “Trust in automated vehicles”. In: *User Experience Design in the Era of Automated Driving*, pp. 29–49.

Izquierdo-Reyes, Javier et al. (2018). “Emotion recognition for semi-autonomous vehicles framework”. In: *International Journal on Interactive Design and Manufacturing (IJIDeM)* 12, pp. 1447–1454.

Jackson, Douglas N, Samuel Messick (1958). “Content and style in personality assessment.” In: *Psychological bulletin* 55.4, p. 243.

Jain, Deepak Kumar et al. (2023). “An automated hyperparameter tuned deep learning model enabled facial emotion recognition for autonomous vehicle drivers”. In: *Image and Vision Computing* 133, p. 104659.

Jaiswal, Dibyanshu et al. (2023). “Gsr based generic stress prediction system”. In: *Adjunct Proceedings of the 2023 ACM International Joint Conference on Pervasive and Ubiquitous Computing & the 2023 ACM International Symposium on Wearable Computing*, pp. 433–438.

Keogh, Eamonn, Chotirat Ann Ratanamahatana (2005). “Exact indexing of dynamic time warping”. In: *Knowledge and information systems* 7, pp. 358–386.

Kim, Kyung Hwan, Seok Won Bang, Sang Ryong Kim (2004). “Emotion recognition system using short-term monitoring of physiological signals”. In: *Medical and biological engineering and computing* 42, pp. 419–427.

Ko, Byoung Chul (2018). “A brief review of facial emotion recognition based on visual information”. In: *sensors* 18.2, p. 401.

Kumar, P Sriram et al. (2023). “A comparative analysis of eda decomposition methods for improved emotion recognition”. In: *Journal of Mechanics in Medicine and Biology* 23.06, p. 2340043.

Kyriakidis, Miltos, Riender Happee, Joost CF de Winter (2015). “Public opinion on automated driving: Results of an international questionnaire among 5000 respondents”. In: *Transportation research part F: traffic psychology and behaviour* 32, pp. 127–140.

Leng, H, Y Lin, LA Zanzi (2007). “An experimental study on physiological parameters toward driver emotion recognition”. In: *Ergonomics and Health Aspects of Work with Computers: International Conference, EHAWC 2007, Held as Part of HCI International 2007, Beijing, China, July 22-27, 2007. Proceedings*. Springer, pp. 237–246.

Li, Shan, Weihong Deng (2019). “Reliable Crowdsourcing and Deep Locality-Preserving Learning for Unconstrained Facial Expression Recognition”. In: *IEEE Transactions on Image Processing* 28.1, pp. 356–370.

— (2020). “Deep facial expression recognition: A survey”. In: *IEEE transactions on affective computing* 13.3, pp. 1195–1215.

Li, Shan, Weihong Deng, JunPing Du (2017). “Reliable Crowdsourcing and Deep Locality-Preserving Learning for Expression Recognition in the Wild”. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 2584–2593.

Li, Wenbo, Yaodong Cui, et al. (2021). “A spontaneous driver emotion facial expression (defe) dataset for intelligent vehicles: Emotions triggered by video-audio clips in driving scenarios”. In: *IEEE Transactions on Affective Computing*.

Lucey, Patrick et al. (2010). “The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression”. In: *2010 ieee computer society conference on computer vision and pattern recognition-workshops*. IEEE, pp. 94–101.

Lutin, Erika et al. (2021). “Feature Extraction for Stress Detection in Electrodermal Activity.” In: *BIOSIGNALS*. Vienna, Austria, pp. 177–185.

Macfarlane, Alexander Blair Stuart (2016). “Modular Electric Automatic Guided Vehicle Suspension-Drive Unit”. PhD thesis. Nelson Mandela Metropolitan University.

Makowski, Dominique et al. (Feb. 2021). “NeuroKit2: A Python toolbox for neurophysiological signal processing”. In: *Behavior Research Methods* 53.4, pp. 1689–1696. DOI: 10.3758/s13428-020-01516-y. URL: https://doi.org/10.3758%2Fs13428-020-01516-y.

Mara, Martina, Kathrin Meyer (2022). “Acceptance of autonomous vehicles: An overview of user-specific, car-specific and contextual determinants”. In: *User experience design in the era of automated driving*, pp. 51–83.

Marangunić, Nikola, Andrina Granić (2015). “Technology acceptance model: a literature review from 1986 to 2013”. In: *Universal access in the information society* 14, pp. 81–95.

Marceddu, Antonio (2020). *Facial Expressions Databases Classifier*. Accessed: 2024-06-24. URL: https://github.com/AntonioMarceddu/Facial_Expressions_Databases_Classifier.

Matsumoto, David (1990). “Cultural similarities and differences in display rules”. In: *Motivation and emotion* 14.3, pp. 195–214.

Memar, Maryam, Amin Mokaribolhassan (2021). “Stress level classification using statistical analysis of skin conductance signal while driving”. In: *SN Applied Sciences* 3.1, p. 64.

Meza-García, B, N Rodríguez-Ibáñez (2021). “Driver’s Emotions Detection with Automotive Systems in Connected and Autonomous Vehicles (CAVs).” In: *CHIRA*, pp. 258–265.

Mollahosseini, Ali, Behzad Hasani, Mohammad H Mahoor (2017). “Affectnet: A database for facial expression, valence, and arousal computing in the wild”. In: *IEEE Transactions on Affective Computing* 10.1, pp. 18–31.

Mou, Luntian, Yiyuan Zhao, et al. (2023). “Driver Emotion Recognition with a Hybrid Attentional Multimodal Fusion Framework”. In: *IEEE Transactions on Affective Computing*.

Mou, Luntian, Chao Zhou, et al. (2021). “Driver stress detection via multimodal fusion using attention-based CNN-LSTM”. In: *Expert Systems with Applications* 173, p. 114693.

Naga, Prameela, Swamy Das Marri, Raiza Borreo (2023). “Facial emotion recognition methods, datasets and technologies: A literature survey”. In: *Materials Today: Proceedings* 80, pp. 2824–2828.

Nedjah, Nadia, Igor Santos, Luiza Mourelle (Apr. 2019). “Sentiment analysis using convolutional neural network via word embeddings”. In: *Evolutionary Intelligence* 15. DOI: 10.1007/s12065-019-00227-4.

Ngwe, Jia Le et al. (2023). “PAtt-Lite: Lightweight Patch and Attention MobileNet for Challenging Facial Expression Recognition”. In: *arXiv preprint arXiv:2306.09626*.

Nordhoff, Sina et al. (2023). “Do driver’s characteristics, system performance, perceived safety, and trust influence how drivers use partial automation? A structural equation modelling analysis”. In: *Frontiers in Psychology* 14, p. 1125031.

O’Mahony, Niall et al. (2020). “Deep learning vs. traditional computer vision”. In: *Advances in Computer Vision: Proceedings of the 2019 Computer Vision Conference (CVC), Volume 1 1*. Springer, pp. 128–144.

Paddeu, Daniela, Graham Parkhurst, Ian Shergold (2020). “Passenger comfort and trust on first-time use of a shared autonomous shuttle vehicle”. In: *Transportation Research Part C: Emerging Technologies* 115, p. 102604.

Pal, Nikhil R, Sankar K Pal (1993). “A review on image segmentation techniques”. In: *Pattern recognition* 26.9, pp. 1277–1294.

Palatinus, Zsolt et al. (2022). “Physiological measurements in social acceptance of self driving technologies”. In: *Scientific reports* 12.1, p. 13312.

Park, Corey, Mehrdad Nojoumian (2022). “Social acceptability of autonomous vehicles: unveiling correlation of passenger trust and emotional response”. In: *International Conference on Human-Computer Interaction*. Springer, pp. 402–415.

Peintner, Jakob et al. (2024). “Driving Behavior Analysis: A Human Factors Perspective on Automated Driving Styles”. In: *2024 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, pp. 3325–3330.

Peng, Chen et al. (2023). “Conceptualising user comfort in automated driving: Findings from an expert group workshop”. In.

Pham, Trong-Dong et al. (2023). “CNN-Based Facial Expression Recognition with Simultaneous Consideration of Inter-Class and Intra-Class Variations”. In: *Sensors* 23.24, p. 9658.

Posada-Quintero, Hugo F, Ki H Chon (2020). “Innovations in electrodermal activity data collection and signal processing: A systematic review”. In: *Sensors* 20.2, p. 479.

Prince, Simon JD (2023). *Understanding deep learning*. MIT press.

Rosenberg, Erika L, Paul Ekman (2020). *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press.

Russell, James A (1980). “A circumplex model of affect.” In: *Journal of personality and social psychology* 39.6, p. 1161.

Safdar, Saba et al. (2009). “Variations of emotional display rules within and across cultures: A comparison between Canada, USA, and Japan.” In: *Canadian Journal of Behavioural Science/Revue canadienne des sciences du comportement* 41.1, p. 1.

Salvador, Stan, Philip Chan (2007). “Toward accurate dynamic time warping in linear time and space”. In: *Intelligent Data Analysis* 11.5, pp. 561–580.

Savran, Arman, Ruben Gur, Ragini Verma (2013). “Automatic detection of emotion valence on faces using consumer depth cameras”. In: *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 75–82.

Selvaraju, Ramprasaath R et al. (2017). “Grad-cam: Visual explanations from deep networks via gradient-based localization”. In: *Proceedings of the IEEE international conference on computer vision*, pp. 618–626.

Serengil, Sefik Ilkin, Alper Ozpinar (2021). “HyperExtended LightFace: A Facial Attribute Analysis Framework”. In: *2021 International Conference on Engineering and Emerging Technologies (ICEET)*. IEEE, pp. 1–4. DOI: 10.1109/ICEET53442.2021.9659697. URL: https://ieeexplore.ieee.org/document/9659697.

Sini, Jacopo, Antonio Costantino Marceddu, Massimo Violante (2020). “Automatic emotion recognition for the calibration of autonomous driving functions”. In: *Electronics* 9.3, p. 518.

Sini, Jacopo, Antonio Costantino Marceddu, Massimo Violante, Riccardo Dessì (2020). “Passengers’ emotions recognition to improve social acceptance of autonomous driving vehicles”. In: *Progresses in Artificial Intelligence and Neural Systems*. Springer, pp. 25–32.

Smith, Steven W et al. (1997). *The scientist and engineer’s guide to digital signal processing*.

Su, Haotian, Johnell Brooks, Yunyi Jia (2023). *Development and Evaluation of Comfort Assessment Approaches for Passengers in Autonomous Vehicles*. Tech. rep. SAE Technical Paper.

Van Vaerenbergh, Yves, Troy D Thomas (2013). “Response styles in survey research: A literature review of antecedents, consequences, and remedies”. In: *International journal of public opinion research* 25.2, pp. 195–217.

Vaswani, Ashish et al. (2017). “Attention is all you need”. In: *Advances in neural information processing systems* 30.

Venkatesh, Viswanath, James YL Thong, Xin Xu (2012). “Consumer acceptance and use of information technology: extending the unified theory of acceptance and use of technology”. In: *MIS quarterly*, pp. 157–178.

Vleugel, Dimitry (Aug. 2024). *Zo populair zijn de robottaxi’s van Waymo: 100.000 betaalde ritten per week*. Accessed: 2024-08-22. URL: https://www.bright.nl/nieuws/1220505/waymo-bereikt-mijlpaal-van-100-000-betaalde-ritten-per-week.html.

Wadud, Zia, Don MacKenzie, Paul Leiby (2016). “Help or hindrance? The travel, energy and carbon impacts of highly automated vehicles”. In: *Transportation Research Part A: Policy and Practice* 86, pp. 1–18.

Wang, Dingyu, Shaocheng Jia, et al. (Dec. 2023). “DERNet: Driver Emotion Recognition Using Onboard Camera”. In: *IEEE Intelligent Transportation Systems Magazine*, pp. 2–17. DOI: 10.1109/MITS.2023.3333882.

Wang, Ke, Yi Lu Murphey, et al. (2019). “Detection of driver stress in real-world driving environment using physiological signals”. In: *2019 IEEE 17th International Conference on Industrial Informatics (INDIN)*. Vol. 1. IEEE, pp. 1807–1814.

Waytz, Adam, Joy Heafner, Nicholas Epley (2014). “The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle”. In: *Journal of experimental social psychology* 52, pp. 113–117.

Weigl, Klemens et al. (2021). “Development of the Questionnaire on the Acceptance of Automated Driving (QAAD): Data-driven models for Level 3 and Level 5 automated driving”. In: *Transportation research part F: traffic psychology and behaviour* 83, pp. 42–59.

Winkel, Ksander N de et al. (2023). “Standards for passenger comfort in automated vehicles: Acceleration and jerk”. In: *Applied Ergonomics* 106, p. 103881.

Wintersberger, Philipp, Andreas Riener, Anna-Katharina Frison (2016). “Automated driving system, male, or female driver: Who’d you prefer? comparative analysis of passengers’ mental conditions, emotional states & qualitative feedback”. In: *Proceedings of the 8th international conference on automotive user interfaces and interactive vehicular applications*, pp. 51–58.

Wu, Jing-Han, Hsing-Wei Lin, Wan-Yu Liu (2020). “Tourists’ environmental vandalism and cognitive dissonance in a National Forest Park”. In: *Urban Forestry & Urban Greening* 55, p. 126845.

Xiao, Huafei et al. (2022). “On-road driver emotion recognition using facial expression”. In: *Applied Sciences* 12.2, p. 807.

Yu, Yong et al. (2019). “A review of recurrent neural networks: LSTM cells and network architectures”. In: *Neural computation* 31.7, pp. 1235–1270.

Zepf, Sebastian et al. (2020). “Driver emotion recognition for intelligent vehicles: A survey”. In: *ACM Computing Surveys (CSUR)* 53.3, pp. 1–30.

Zhang, Saining et al. (2023). “A dual-direction attention mixed feature network for facial expression recognition”. In: *Electronics* 12.17, p. 3595.

A

Source Code

```python
"""
Record the webcam footage. Will store the frames as video and a dataframe with a
    corresponding to each frame.
"""

import time
import os
import cv2
import pandas as pd
import threading
import sys

name = "participant_36" # number of the participant
style = "random_round" # set or random round

exit_flag = False

def check_input():
    global exit_flag
    while True:
        user_input = input()
        if user_input == 'quit':
            exit_flag = True
            break


input_thread = threading.Thread(target=check_input)
input_thread.start()

print("Type quit to stop recording")


path_name = name + "_" + style
path = os.path.dirname(os.path.abspath(__file__))
raw_path = os.path.join(path, f'{path_name}.avi')
data_path = os.path.join(path, f'{path_name}.csv')

fourcc = cv2.VideoWriter_fourcc(*'XVID')
raw_footage = cv2.VideoWriter(raw_path, fourcc, 30.0, (640, 480))

df_timestamps = pd.DataFrame()

cap = cv2.VideoCapture(0)
if not cap.isOpened():
    print("Error: could not open video")
    exit()

frame_count = 1
timestamp_list = []
frame_list = []
```

```

while True:
    if exit_flag:
        break

    ret, frame = cap.read()
    raw_footage.write(frame)

    timestamp_list.append(time.time_ns())
    frame_list.append(frame_count)
    frame_count += 1

cap.release()
raw_footage.release()
cv2.destroyAllWindows()

df_timestamps['frame'] = frame_list
df_timestamps['timestamp'] = timestamp_list
df_timestamps.to_csv(data_path, index=False)

input_thread.join()
```

```
import cv2
import os
from retinaface import RetinaFace
import numpy as np
from tqdm import tqdm
import argparse

def align_face(img, landmark):
    src = np.array([
        [38.2946, 51.6963],  # Left eye
        [73.5318, 51.5014],  # Right eye
        [56.0252, 71.7366],  # Nose tip
        [41.5493, 92.3655],  # Left mouth corner
        [70.7299, 92.2041]   # Right mouth corner
    ], dtype=np.float32)

    dst = np.array([
        landmark["right_eye"],   # Right eye from the person's perspective
        landmark["left_eye"],    # Left eye from the person's perspective
        landmark["nose"],
        landmark["mouth_right"], # Right mouth corner from the person's perspective
        landmark["mouth_left"]   # Left mouth corner from the person's perspective
    ], dtype=np.float32)

    tform = cv2.estimateAffinePartial2D(dst, src)[0]
    face_img = cv2.warpAffine(img, tform, (112, 112), borderValue=0.0)

    return face_img

def extract_frames_and_save(video_path, output_folder):
    cap = cv2.VideoCapture(video_path)
    frame_count = int(cap.get(cv2.CAP_PROP_FRAME_COUNT))

    if not os.path.exists(output_folder):
        os.makedirs(output_folder)

    for i in tqdm(range(frame_count), desc="Processing␣frames"):
        ret, frame = cap.read()
        if not ret:
            break

        faces = RetinaFace.detect_faces(frame)
        if faces is not None:
            for key in faces.keys():
                face = faces[key]
                landmarks = face["landmarks"]
                aligned_face = align_face(frame, landmarks)

                frame_output_folder = os.path.join(output_folder, 'aligned_faces')
```

```
                if not os.path.exists(frame_output_folder):
                    os.makedirs(frame_output_folder)

                output_path = os.path.join(frame_output_folder, f'frame_{i:04d}.jpg')
                cv2.imwrite(output_path, aligned_face)

    cap.release()

if __name__ == '__main__':
    parser = argparse.ArgumentParser()
    parser.add_argument('--video_path', type=str, required=True, help='Path to the input video file.')
    parser.add_argument('--output_folder', type=str, required=True, help='Path to the output folder to save aligned faces.')
    args = parser.parse_args()

    extract_frames_and_save(args.video_path, args.output_folder)
```

B

Questionnaire

Instructions

Rate the following 8 statements from 1 (do not agree at all) to 5 (completely agree). When an automated vehicle is mentioned, automation level SAE4+ is meant. This means that the vehicle can drive fully automated, without human supervision. After every question you have the option to provide comments/context to your given answer. If anything is unclear, please ask the researcher.

Questionnaire

1. **I trust the technology behind automated vehicles to work correctly.**
   Not at all 1 2 3 4 5 Completely
   Why?

2. **I think riding in an automated vehicle would be comfortable.**
   Not at all 1 2 3 4 5 Completely
   Why?

3. **I would feel relaxed while being in an automated vehicle.**
   Not at all 1 2 3 4 5 Completely
   Why?

4. **I believe automated vehicles could reduce traffic accidents.**
   Not at all 1 2 3 4 5 Completely
   Why?

5. **I would rather trust a fully automated vehicle than today's drivers.** **Weigl et al. 2021**
   Not at all 1 2 3 4 5 Completely
   Why?

6. **I would be concerned about safety, if I was driven by a fully automated vehicle.** **Weigl et al. 2021**
   Not at all 1 2 3 4 5 Completely
   Why?

7. **I would not engage in non-driving related activities, but monitor the driving system. Weigl et al. 2021**
   Not at all 1 2 3 4 5 Completely
   Why?

8. **I would use a fully automated vehicle if they are available.Weigl et al. 2021**
   Not at all 1 2 3 4 5 Completely
   Why?

9. **Feel free to share any additional thoughts or comments on fully automated vehicles:**

C

Consent Form

Project Overview

The field of automated driving is developing very fast; however, research has shown that user acceptance of these features is falling behind and requires more attention. The comfort of the user in terms of trust and perceived safety therefore needs to be assessed. The desire is to find an objective measure for this so that the automated vehicle can monitor the state of the passengers.

The goal of this experiment is to assess how facial emotion detection can be used to assess comfort in automated vehicles. For this experiment, we are looking at SAE level 4+ automation. This means you as a human are not driving the vehicle even when you are seated in the “driver’s” seat. You are not required to take over and pedals and/or a steering wheel might not be installed in the vehicle. We want to ask you to keep this scenario in mind during the experiment.

NOTE: This research is not aimed at the area of motion sickness; however, it might occur. If this occurs, please report this to the researcher. It is possible to take a break in the experiment. The participant is allowed to quit the experiment at any moment.

Declaration of Consent

Controller and categories of personal data

Siemens Digital Industries Software NV, Interleuvenlaan 68, 3001 Leuven, Belgium (”Siemens”) collects and processes the following personal data about you:

- The answers to the survey questions
- Contact details to obtain results in case of interest by the participant.

Special categories of personal data

You understand and agree that Siemens collects and processes the following personal data that enjoy special protection under applicable law (so-called “special categories of personal data”):

- **Biometric data:**
  - Video recording of passenger pupils using eye-tracking glasses and the face using a dashboard camera.
  - The video recordings are not studied in detail but only derived characteristics such as eye blinking rate, eye fixation time, or facial emotion classification are used for further processing.
  - The processed parameters on the pupil recordings are computed by uploading these to a cloud environment owned by the eye tracking device company. However, the company does not use this data further and deletion of the uploaded data can be requested at any time.
  - The processed parameters on the dashboard camera recordings are computed in house.
  - Audio recording of voice. This data is only used to check after the measurement if the passenger indicated verbally a change in its comfort level.
  - Heart rate measurement using ECG pads. The processed parameters are computed in house.

– Skin conductance measurement using Velcro finger straps. The processed parameters are computed in house.

ALL data related to the pupils' recordings is pseudonymized; that is, it is labelled with a certain passenger number and not linked to the personal details of this passenger.

Purpose of the processing

Your personal data are collected and processed for the following purposes:

- To conduct the experiment with focus on the thesis.
- To process the recordings and survey results for further Siemens research projects related to the development of the prediction models.

Retention Periods

After completion of a set period, the dataset is deleted in an audit-proof manner. The clear data will be deleted if there is a specific request by the participant according to the rights described under the GDPR. In other cases, the data will be retained in any case for 8-10 years beyond the duration of a research project (typically 3 years) or a commercial project (typically 1 year). Algorithm development should be done on the clear data as algorithm development done on processed (e.g., anonymized) data will result in lower quality algorithms, which conflicts with the overall objective – improving the safety of autonomous systems. Hence, clear data is kept as long as necessary to develop the algorithms and methodologies and integrate them in Siemens' products or services, while keeping the possibility to continuously refine and improve them after that, as well as use them for scientific purposes, also under the provision that the deletion of the data is not likely to prevent or seriously affect the achievement of the overall objective.

Recipients

For the purposes outlined above, your personal data may be transferred to and processed by the following recipients (each a "Recipient"):

- Siemens Industry Software Netherlands B.V (BIC 1, 5657 BX Eindhoven, Netherlands)
- Pupil Labs GmbH (Gustav-Müller-Str. 7, 10829 Berlin; only relevant for the recording of the pupils which gets uploaded to the cloud offering of this vendor)

Consent

I hereby consent to the collection and processing of my personal data as described above. I understand that the provision of my consent is voluntary and that I have the right to withdraw my data protection consent with future effect at any time by contacting m In case you with-draw your consent, Siemens may only further process your personal data where there is another legal ground for the processing.

Signed by: _

Insert Full Name: _

Date Signature: _

Additional consent for publication of image(s) in thesis

We would appreciate your permission for images with your face being used in possible publications of this thesis only. Even when permission is given, the researcher will always check for the specific image in question before using it in the publication. Please select your choice below by circling the relevant option.

☐ No image material of me can be published in the thesis.

☐ Images can be used but no recognizable full-face images. E.g., eyes blocked out.
E.g., zoomed in parts of the face.

☐ Full-face images can be used in the publication of the thesis.

Further Information

Right of access to and rectification or erasure of personal data, restriction of processing, right to object to processing, and right to data portability

Under applicable data protection law, you may - provided that the respective legal preconditions are met - have the right to:

- Obtain from Siemens confirmation as to whether or not personal data concerning you are being processed and where that is the case, access to the personal data;
- Obtain from Siemens the rectification of inaccurate personal data concerning you;
- Obtain from Siemens the erasure of your personal data;
- Obtain from Siemens restriction of processing regarding your personal data;
- Obtain from Siemens a copy of personal data concerning you which you actively provided in a structured, commonly used, and machine-readable format and to request from Siemens that we transmit those data to another recipient selected by you; and
- Object on grounds relating to your particular situation to processing of personal data concerning you.

Data Privacy Contact

The Siemens Data Protection Organization provides support with any data privacy-related questions, comments, concerns, or complaints or in case you wish to exercise any of your data privacy-related rights. The Siemens Data Privacy Organization may be contacted at: dataprotection@siemens.com. The Siemens Data Privacy Organization will always use best efforts to address and settle any requests or complaints you bring to its attention. Besides contacting the Siemens Data Privacy Organization, you always have the right to approach the competent data protection authority with your request or complaint.

Further information concerning the processing of my personal data can be found in the data privacy notice of Siemens Digital Industries Software.

D

FER data plots

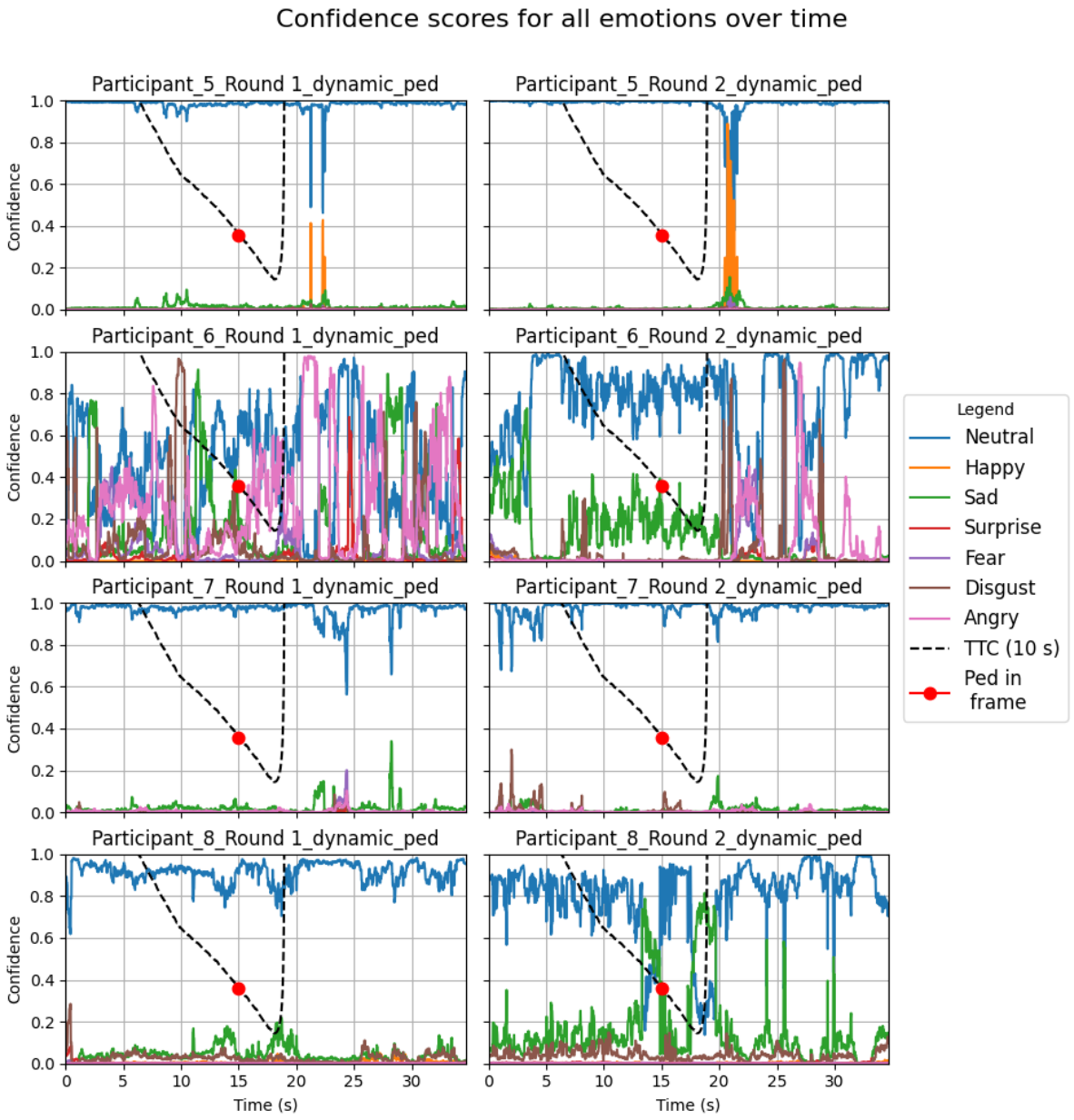


Figure D.1: All emotions confidence scores during the dynamic scenario with the pedestrian in the scene, after applying the moving average filter, for participants 5 through 8.

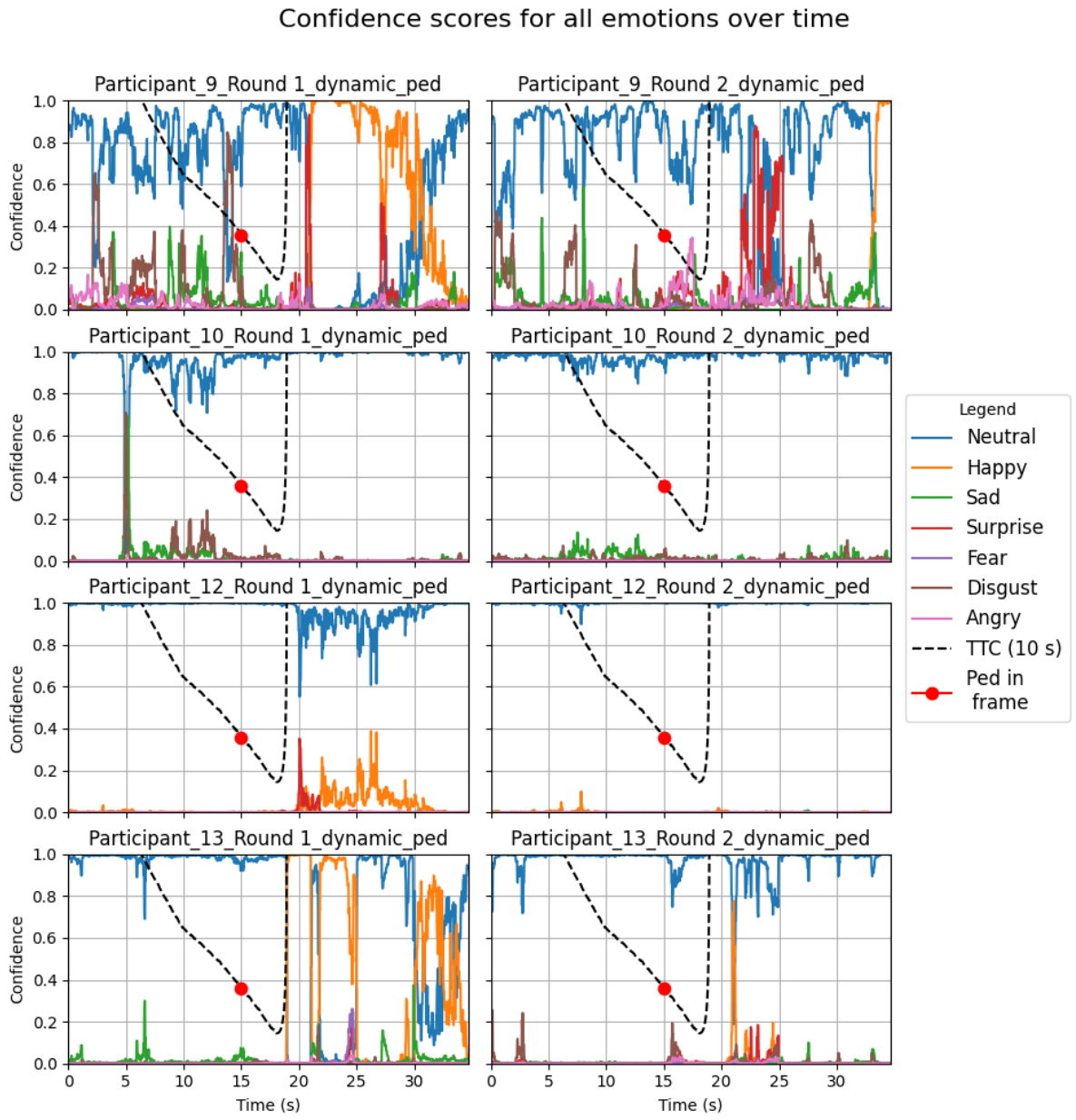


Figure D.2: All emotions confidence scores during the dynamic scenario with the pedestrian in the scene, after applying the moving average filter, for participants 9 through 13, excluding 11.

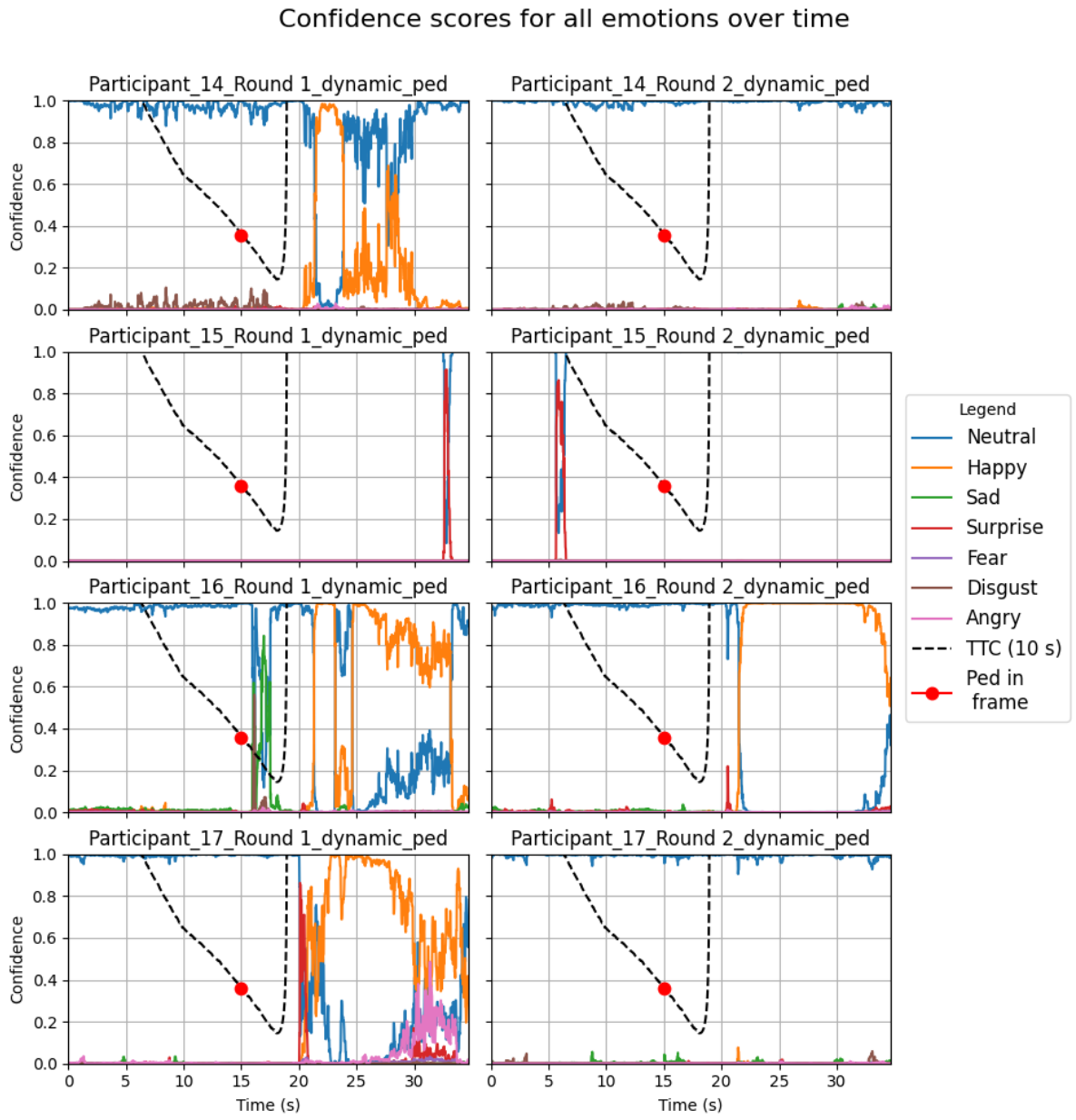


Figure D.3: All emotions confidence scores during the dynamic scenario with the pedestrian in the scene, after applying the moving average filter, for participants 14 through 17.

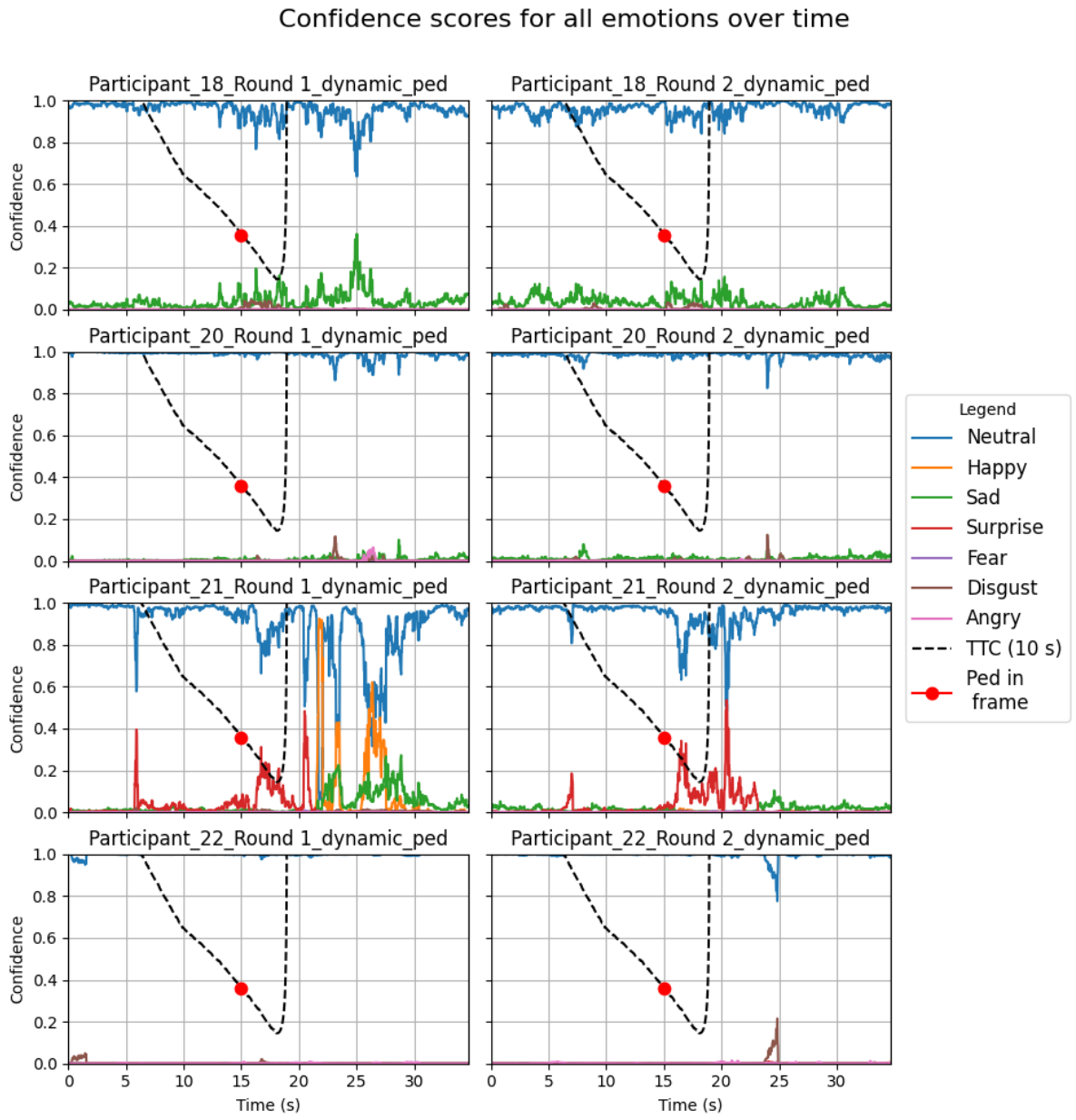


Figure D.4: All emotions confidence scores during the dynamic scenario with the pedestrian in the scene, after applying the moving average filter, for participants 18 through 22, excluding 19.

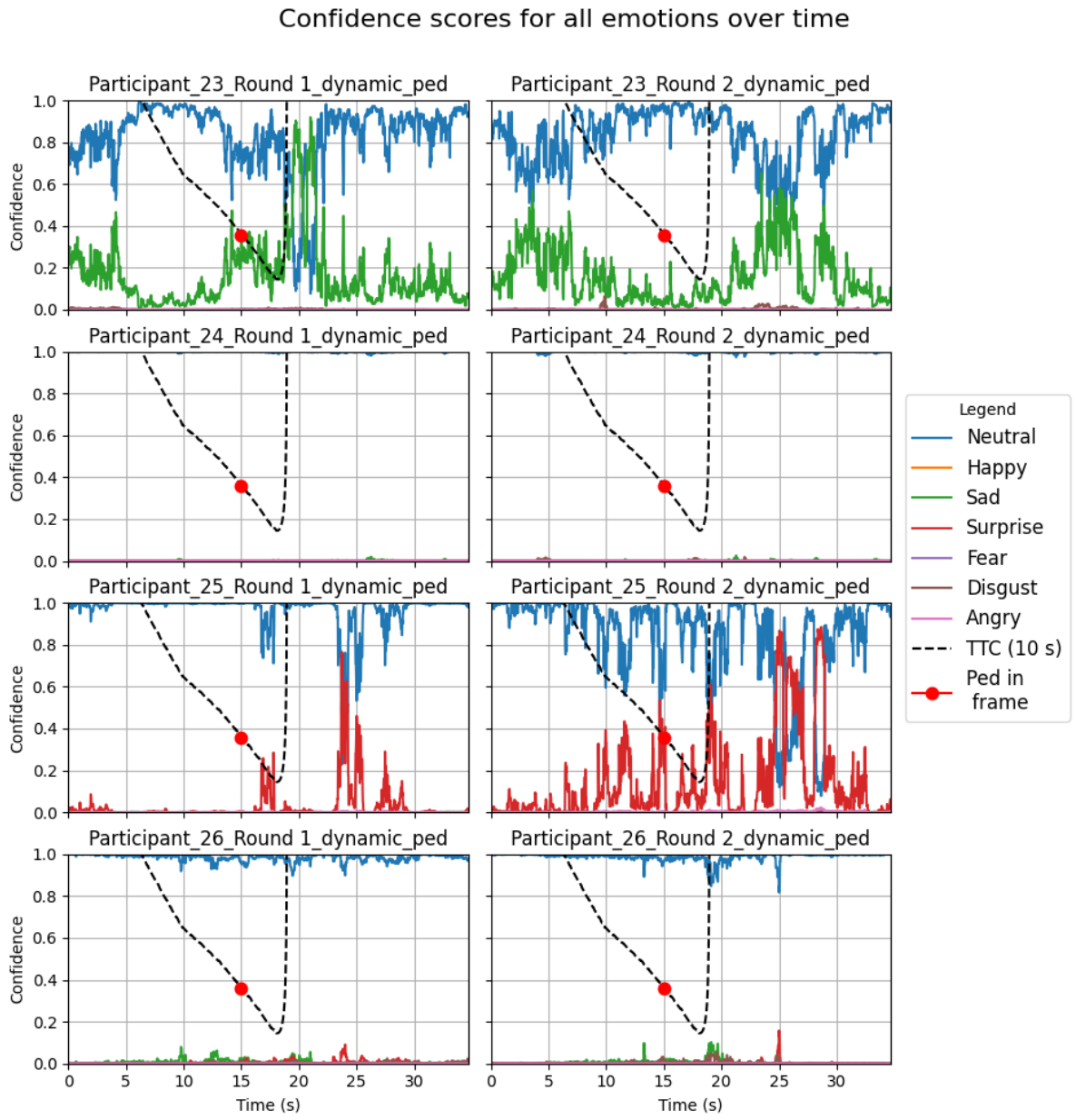


Figure D.5: All emotions confidence scores during the dynamic scenario with the pedestrian in the scene, after applying the moving average filter, for participants 23 through 26.

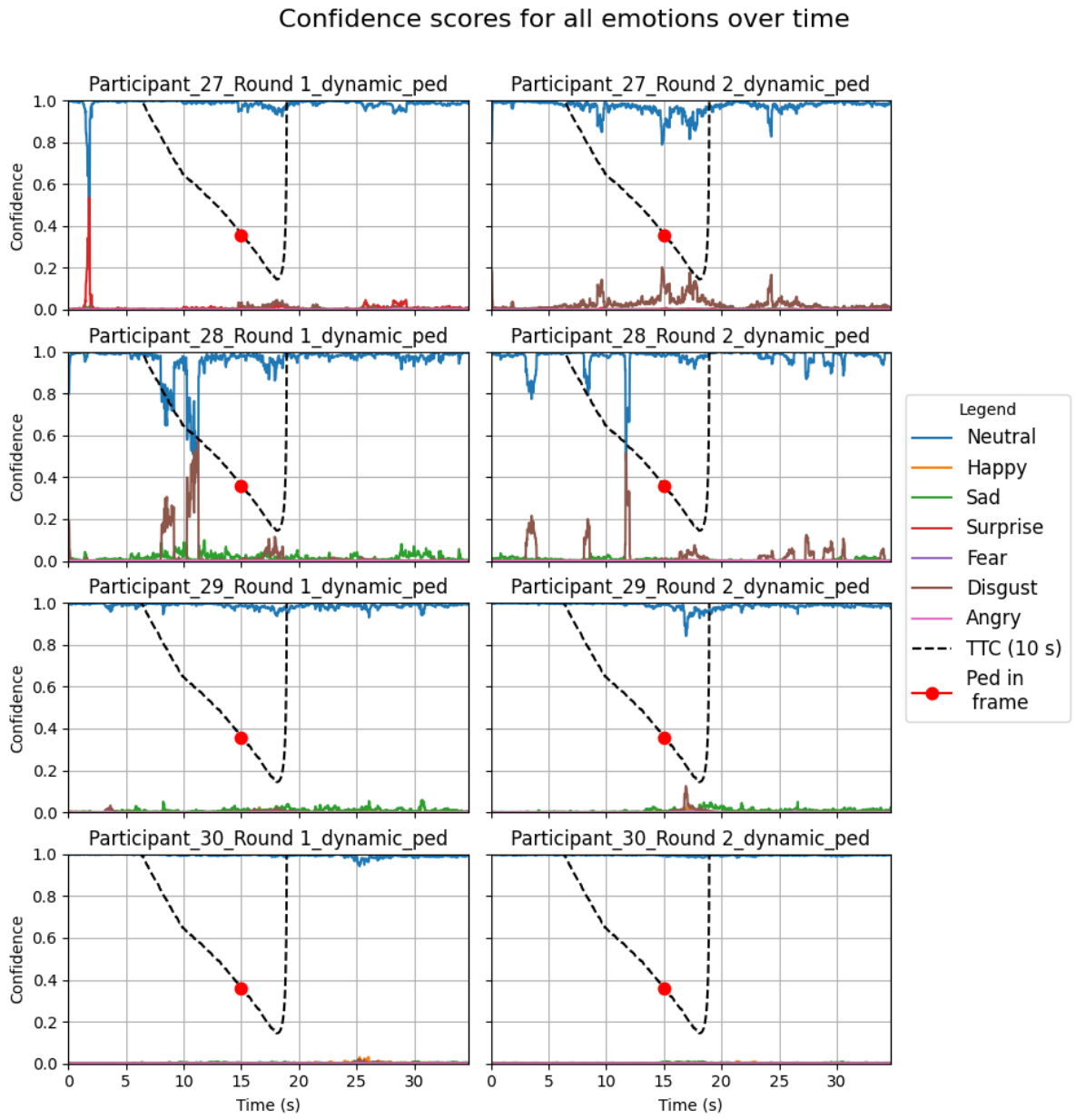


Figure D.6: All emotions confidence scores during the dynamic scenario with the pedestrian in the scene, after applying the moving average filter, for participants 27 through 30.

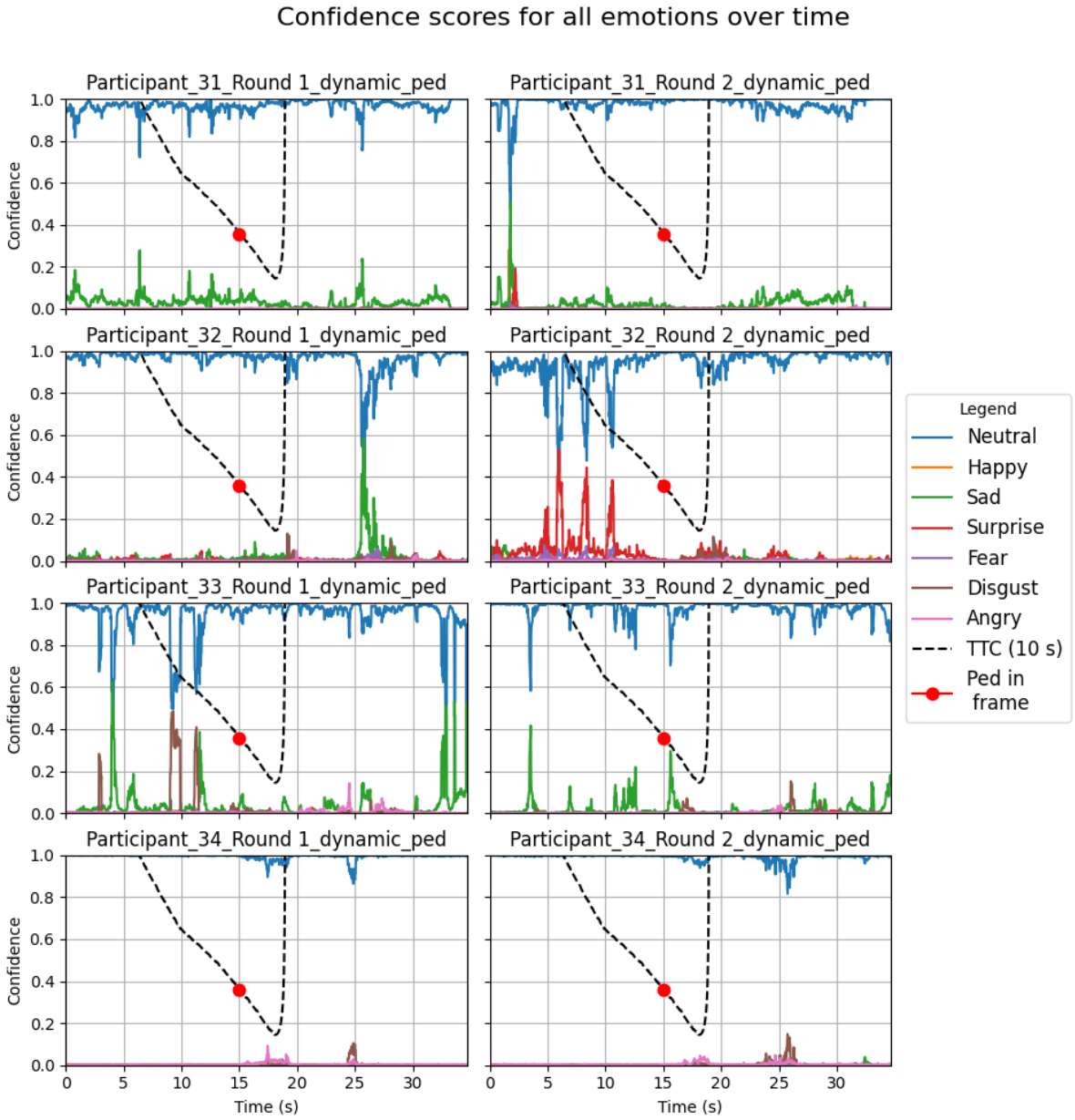


Figure D.7: All emotions confidence scores during the dynamic scenario with the pedestrian in the scene, after applying the moving average filter, for participants 31 through 34.

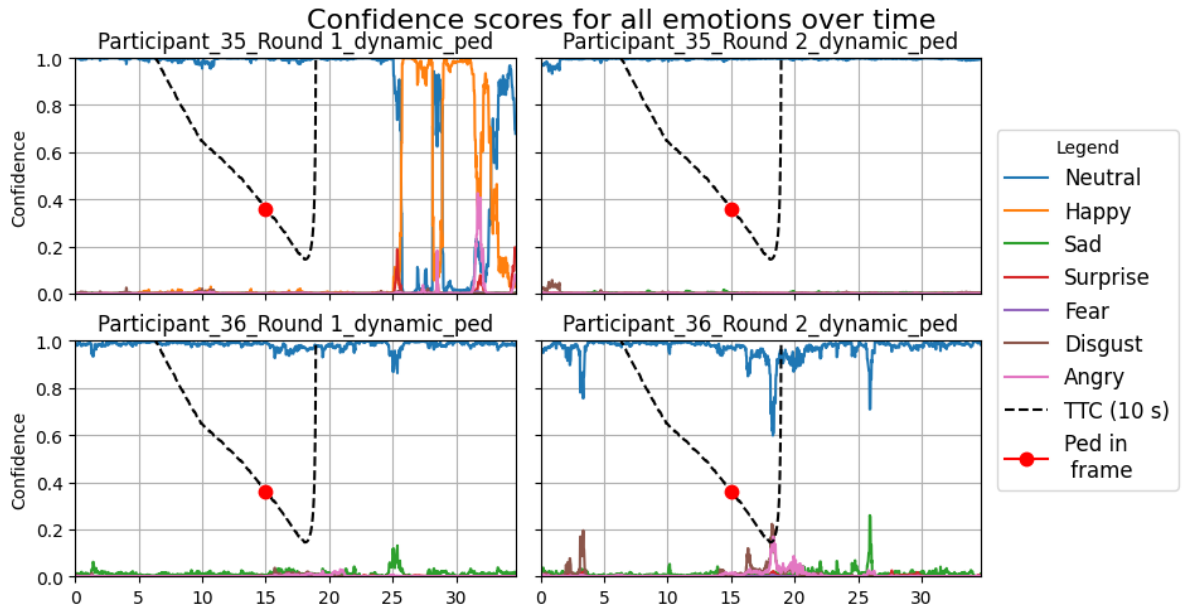


Figure D.8: All emotions confidence scores during the dynamic scenario with the pedestrian in the scene, after applying the moving average filter, for participants 35 and 36.

E

Manuals

E.1. NeXus-10 MKII

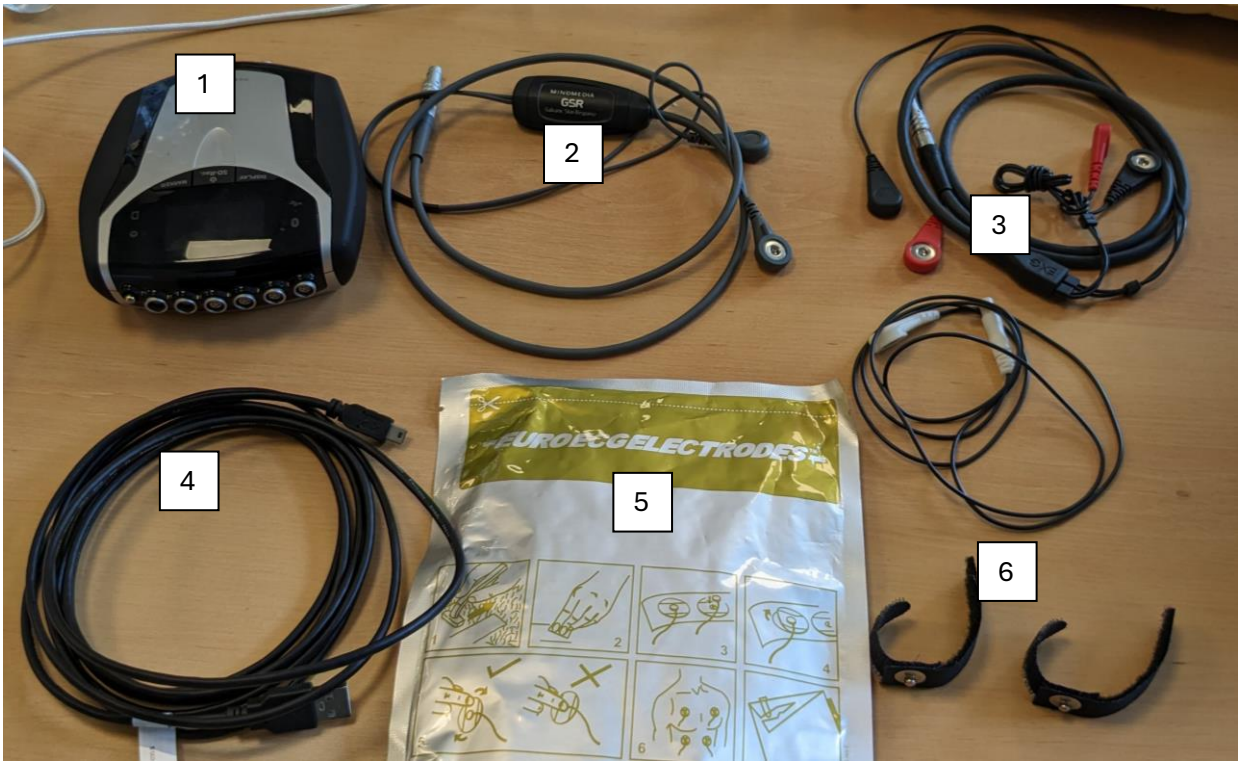
The PDF version of the manual was included in overleaf. It starts on the next page.

NeXus-10 MKII

A device developed by Mindmedia to measure different physiological sensors simultaneously. Visit https://mindmedia.freshdesk.com/support/solutions/folders/36000184533 to find their user manuals. Before starting, follow the BioTrace+ software installation guide (First chapter of the BioTrace+ manual). You also have to make an account. When verifying this account in your email it might come up as 'failed', but this still appeared to work.

Hardware setup

For measuring heart rate (variability) and galvanic skin response, the following equipment is required:



1. The NeXus-10 device can be considered as a "mini-scadas". Works wireless (Bluetooth) and via cable.
2. The GSR sensor, is to be plugged into port E of the NeXus-10 device.
3. The heart rate sensors.
    a. The red and black is the EXG sensors. Only red-2 and black-2 are used for measuring ECG. Is to be connected to port A&B of the NeXus-10.
    b. The white sensor is the Ground Sensor, to be connected to the Gnd port of the NeXus-10.
4. The USB 2.0 connection cable, from type A (to be inserted in the laptop) to type miniB (connection to the NeXus-10). These cables are rarely used anymore, as you will not easily find one.
5. One-time-use ECG electrodes. These "stickers" are used for the heart rate. A pack contains 50 stickers, for each measurement 3 are used.
6. Re-usable GSR Velcro electrodes.

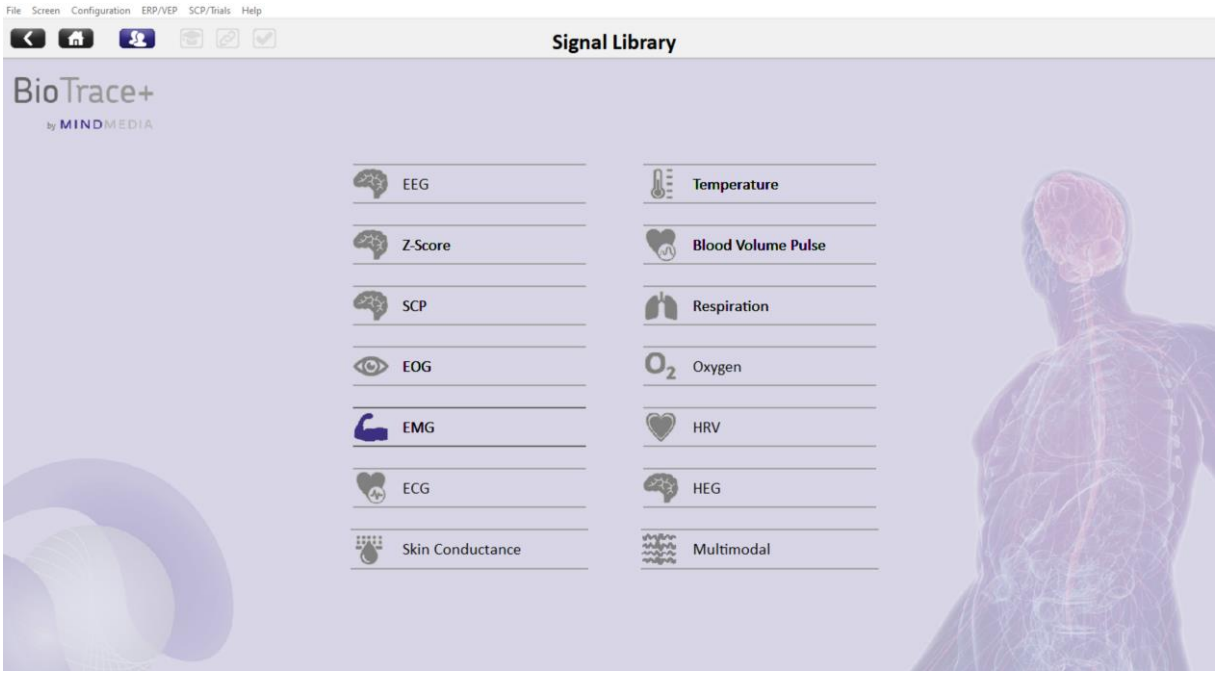
Make sure the NeXus-10 is charged. The battery can be taken out from below, pushing the black slider at the bottom up.

The cable from the adapter to the power socket is not included. This is the same cable as the laptop chargers use, so it should not be hard to find in the office.

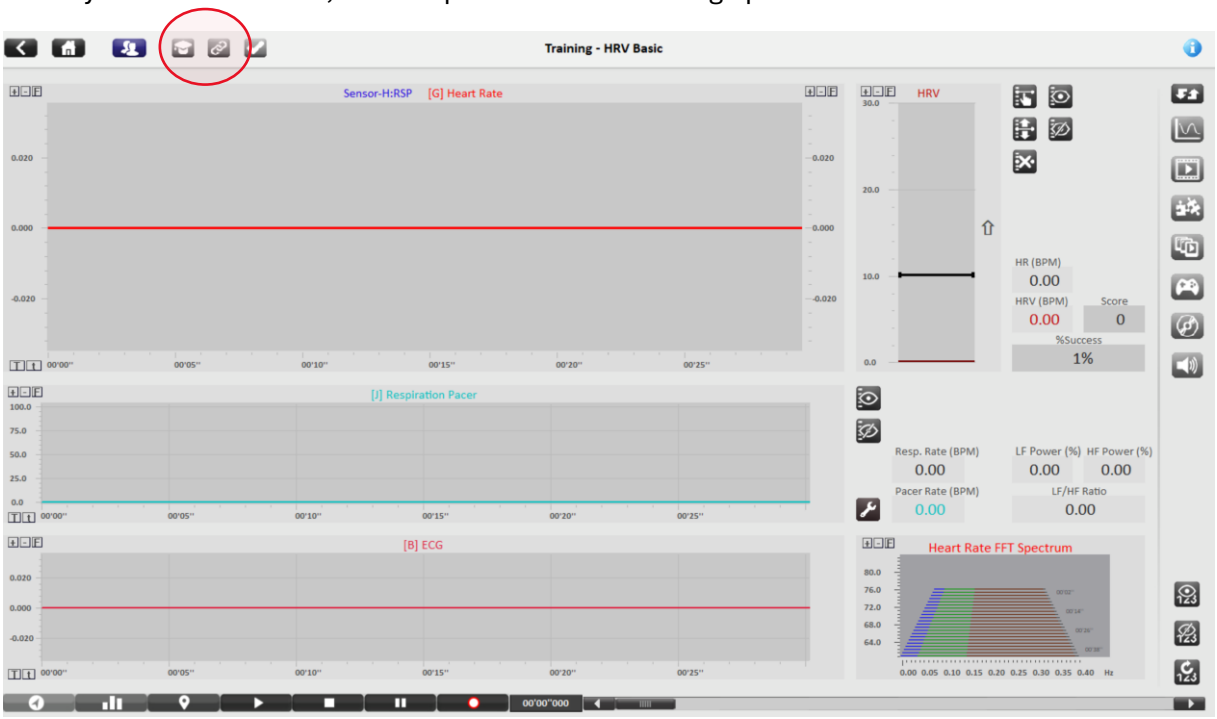
After connecting the sensors, you can turn on the NeXus-10. You will see the blue lights turn on, and the device will check if a memory card is inserted. It will also immediately turn on Bluetooth, but if the USB cable is connected to the laptop, it will stop Bluetooth and say that the wired connection was found. The Bluetooth connection did not seem to work with Windows 11, and in general, to avoid delays, it is recommended to use the cable.

The BioTrace+ software

Unfortunately, the NeXus-10 only works with this software developed by Mindmedia. After installing, you can start the software (note that it will play a sound when starting and closing the software, so if you are in a quiet office you might want to turn off your volume). In the home screen you can navigate to Signal Library. This brings you to an overview of the different possible signals:



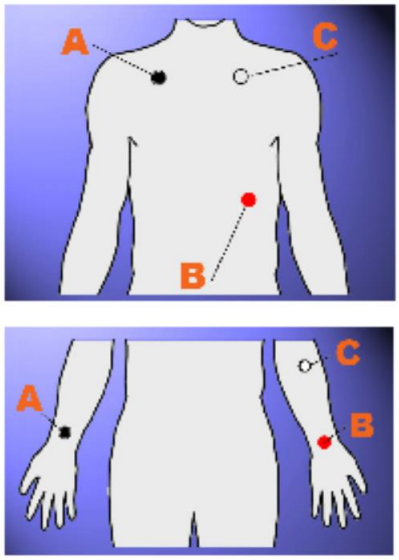
If you select a signal (for our application ECG/HRV and Skin Conductance), you will get a menu of different screens to choose from. (ECG-)HRV Basic or SC Basic are used for our application. When you select a screen, at the top are some interesting options to look at:



Clicking the academic cap icon will bring you to a page with some information about the signal and what it is generally used for. The chain link icon shows a page that tells you how to connect the sensors (to which port of the NeXus-10, and where on the body).

For GSR it is relatively easy. The sensor should be plugged in port E of the NeXus-10, and the Velcro straps should be attached to two fingertips of the same hand. This would preferably be the non-dominant hand, as this is likely the hand that will move the least. Movement can influence the data, so make sure the Velcro straps are tight enough to keep the sensors in place, of course without harming the participant.

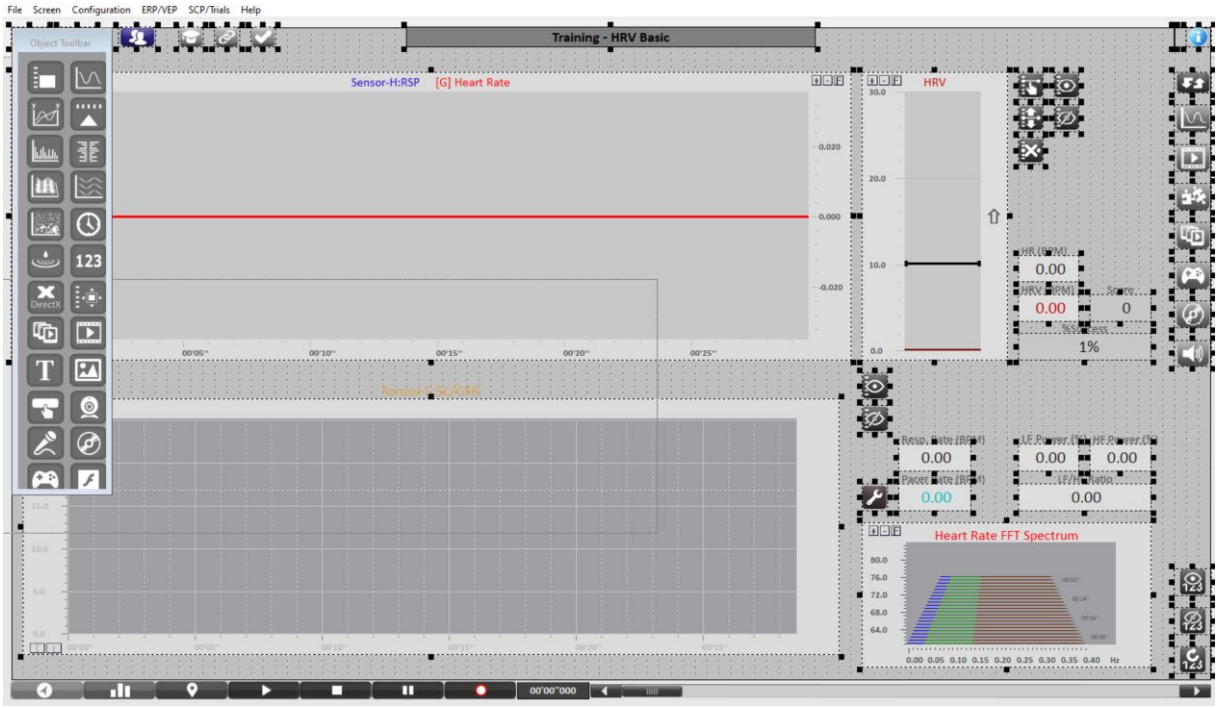
For the heart rate each color has a specific location on the torso, as shown in the picture here. The locations on the arms can be easier to apply, but also gives less accurate measurements.



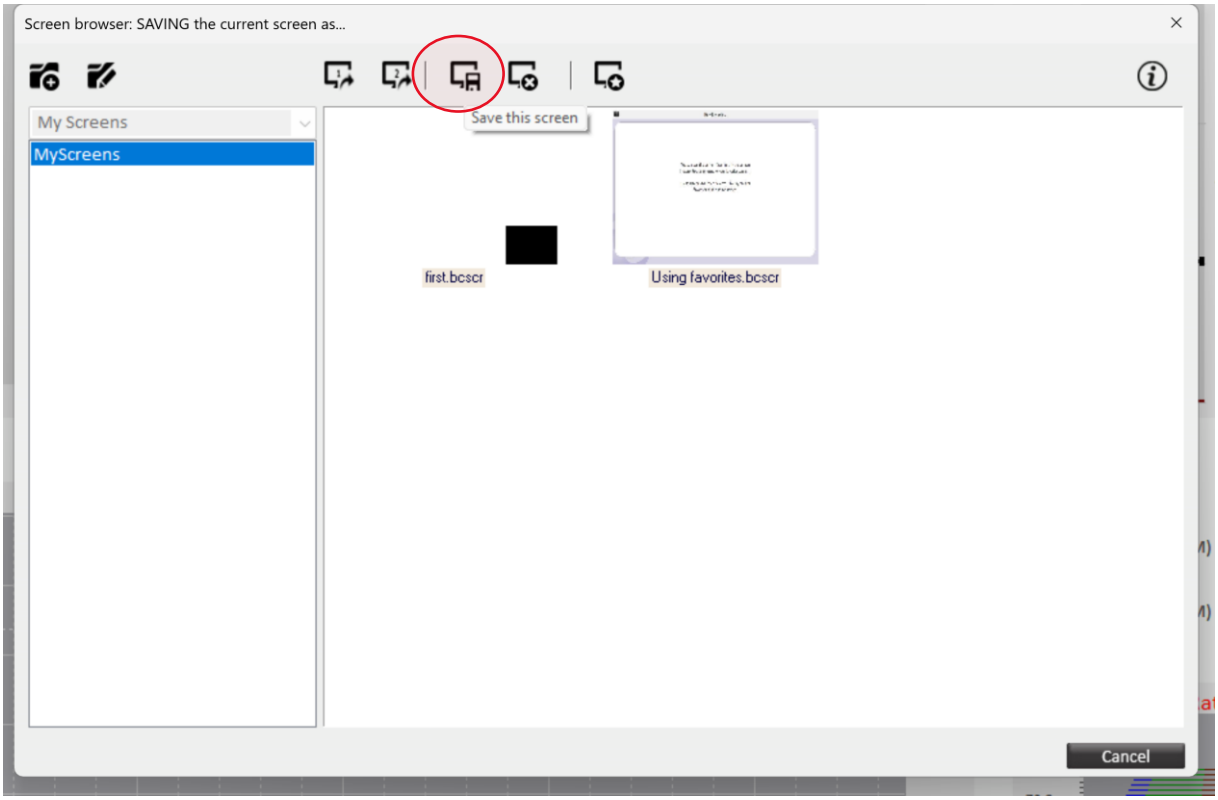
A: Black electrode (negative input)
B: Red electrode (positive input)
C: Ground electrode

Performing measurements

To start measuring, you need to create a screen. The easiest way to do this is to go to the screen of one of the sensors, e.g., HRV-Basic, and click the “e” on your keyboard. You can now edit the screen. You can delete graphs and plots that you do not need and add ones that you would like to see. When you are done again click “e” on the keyboard again to go out of editing mode.

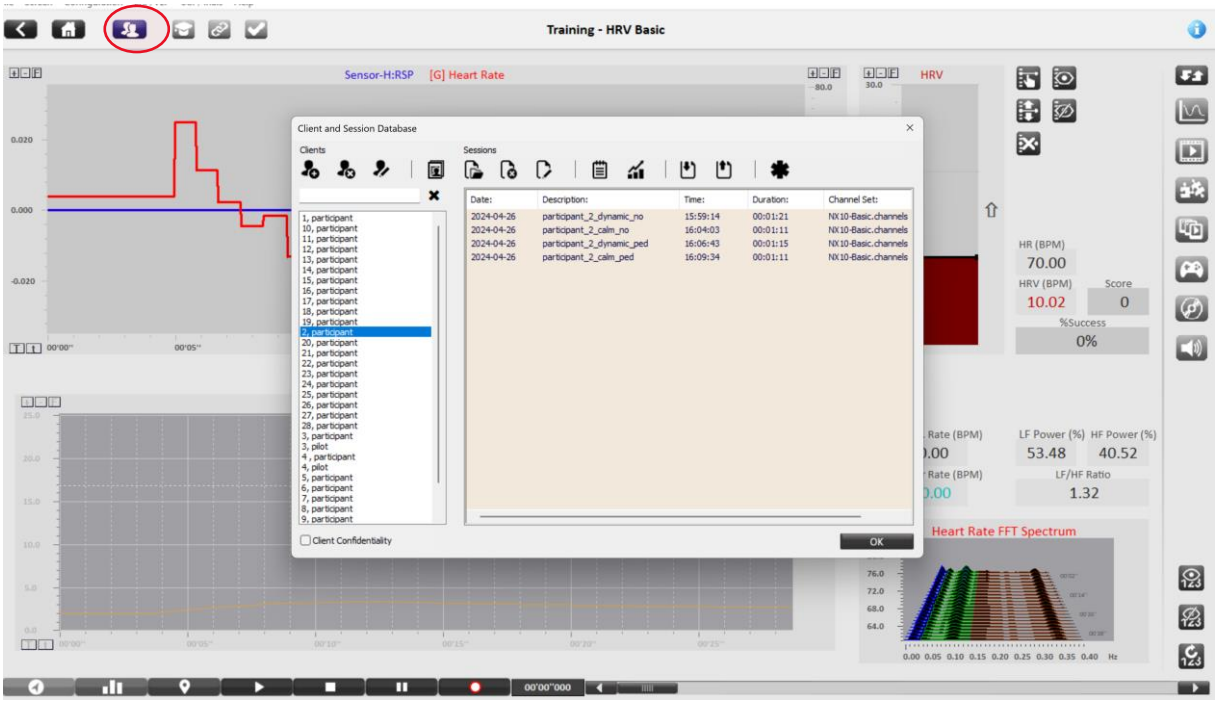


After you are done, go to file→save screen as… and the following window pops up:



Click save this screen and give the screen a name. Now, when you are in the home screen of the software, you can click the L on your keyboard to load a screen.

You also want to create clients. This is done by clicking the purple button with the icons of people in it



This brings you to the client database, where you can see the clients, create new ones, and see al the recorded sessions per client.

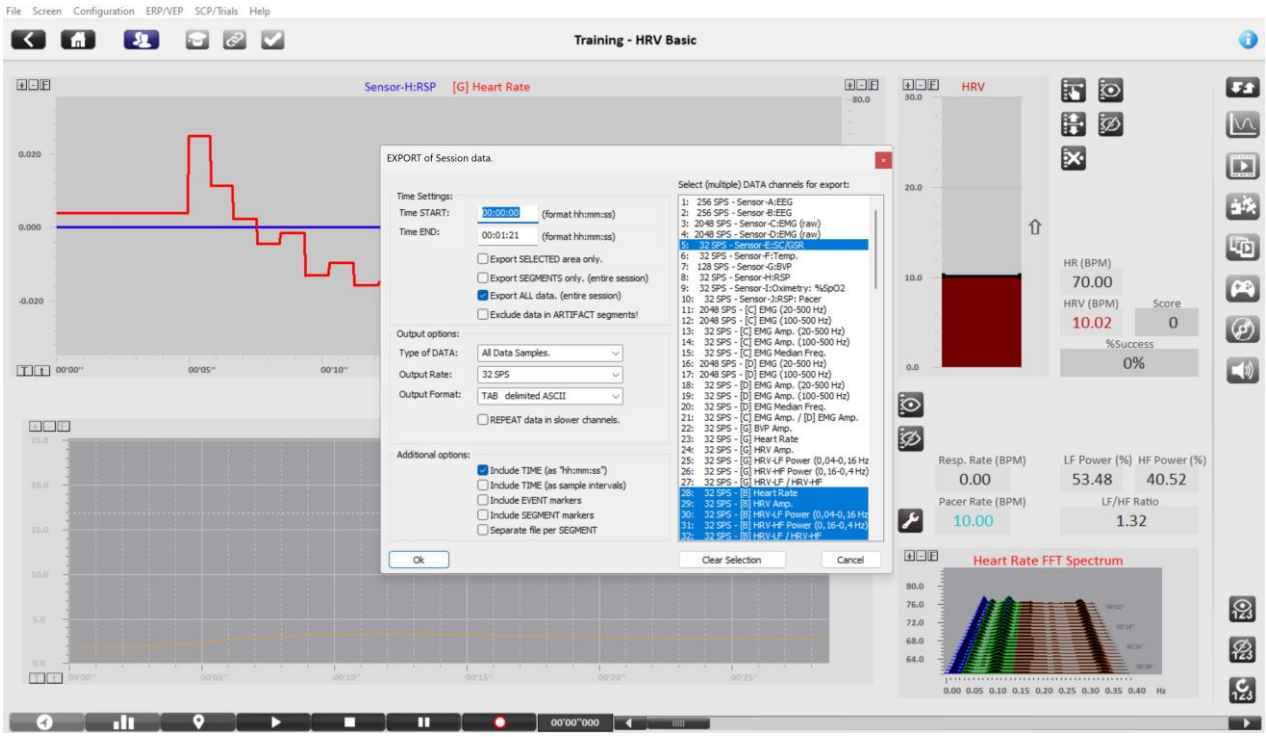
When you want to start a recording, click the red record button on the bottom of your screen. A window will pop up where you select the client that you want to record for. After that it will take a few seconds to get the system ready, and then you can click “START RECORDING”

While the recording is going, you can watch the signals live you the screen.

To stop the recording, you click the button with the square stop icon. It will ask to save the data, and after confirming, you can give the session a name.

Exporting and processing the data

In the client database, select the session that you want to export by double clicking it. Then go to file→export session data, the following window pops up:



In the right half, select the data channels to export. For our application select number 5 (GSR) and number 28-32 (HR(V)) signals. You can also check that their names match with what you want.

You can select which timestamps you want to include, unfortunately they are always relative to the start, and only in seconds even though the data is in 32Hz.

Click OK, navigate to the location where you want the data stored, and give the file a name.

To get timestamped dataframes from these txt files, we use a python script. The script can be found on the next pages. This script will go through a directory, and convert all files to data frames with UNIX timestamps.

Note: These timestamps will have a fault margin of 1 second, as that is the precision with which the original data is given. Make sure to check and adapt the file paths to match your directory names. Also, in this script, 2 hours were subtracted to compensate for the different time zone that the system was working in. This can vary depending on the system you are working on!

```
import pandas as pd
import os
from datetime import datetime, timedelta


def parse_file(file_path):
    with open(file_path, 'r') as file:
        lines = file.readlines()

    # Check if the first line is as expected
    if not lines[0].strip().startswith('RAW Data export file (tab
separated)'):
        raise ValueError(f"File {file_path} does not start with expected
header")

    # Determine the start of the actual data and extract metadata
    data_start_index = None
    headers = None
    start_time = None
    date_str = None
    for i, line in enumerate(lines):
        if 'TIME\tSensor-E:SC/GSR' in line:
            data_start_index = i + 1
            headers = line.strip().split('\t')
        if line.startswith("Time:"):
            start_time_str = line.split('\t')[1].strip()
        if line.startswith("Date:"):
            date_str = line.split('\t')[1].strip()

    if data_start_index is None or start_time_str is None or date_str is None:
        raise ValueError("Data start point, start time, or date not found in
the file")

    # Combine date and time for the full datetime
    start_datetime = datetime.strptime(f"{date_str} {start_time_str}", '%Y-%m-
%d %H:%M:%S')

    # Filter and prepare data lines
    valid_data_lines = []
    for line in lines[data_start_index:]:
        if line.strip() and not line.startswith('<end of exported RAW data>'):
            valid_data_lines.append(line)

    # Read the data into a DataFrame
    from io import StringIO
    data_string = ''.join(valid_data_lines)
    data = pd.read_csv(StringIO(data_string), sep='\t', names=headers)
```

```
    # Adjust TIME column to more precise timing and add Unix timestamp
    sample_interval = timedelta(seconds=1/32)  # Sample interval for 32 Hz
    data['TIME'] = [start_datetime + i * sample_interval for i in
range(len(data))]
    data['UNIX TIMESTAMP'] = data['TIME'].apply(lambda x: x.timestamp()) -
2*60*60 # subtract two hours to get the correct time zone
    # data['UNIX TIMESTAMP'] -= 37  # Subtract 37 seconds to align with
platform

    return data

def process_directory(directory_path, destination_directory):
    files = [os.path.join(directory_path, f) for f in
os.listdir(directory_path) if os.path.isfile(os.path.join(directory_path, f))
and f.endswith('.txt')]

    for file_path in files:
        try:
            dataframe = parse_file(file_path)
            # Save each DataFrame to a separate file
            output_path = os.path.join(destination_directory,
os.path.basename(file_path).replace('.txt', '_processed.txt'))
            dataframe.to_csv(output_path, sep='\t', index=False)
            print(f"Processed and saved: {output_path}")
        except Exception as e:
            print(f"Failed to process {file_path}: {e}")

# Specify the directory path here
participants = list(range(5, 37))
for participant in participants:
    participant = f'participant_{participant}'
    directory_path = f'data/{participant}/heartrate_GSR/raw_data'
    destination_directory = f'data/{participant}/heartrate_GSR/processed_data'
    process_directory(directory_path, destination_directory)
```

E.2. Pupil Labs Invisible eye tracker

The PDF version of the manual was included in overleaf. It starts on the next page.

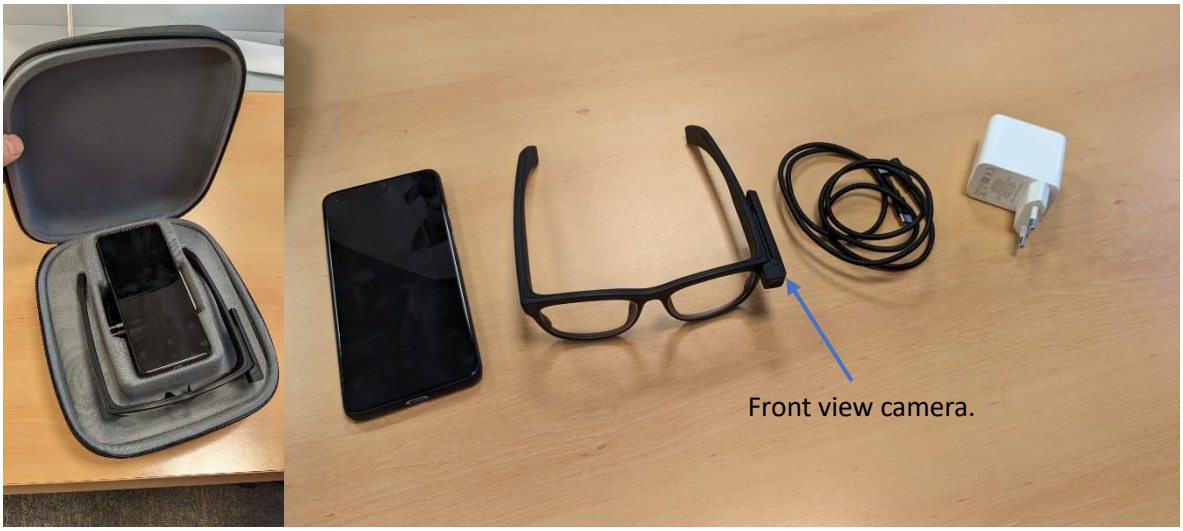
Pupil Invisible glasses

These glasses are developed by Pupil Labs. They have a lot of information and instructions on how to use it accessible on their documentation pages: https://docs.pupil-labs.com/invisible/

Hardware

The case should contain the following:

- Smartphone (OnePlus 8T)
- Warp charge 65 adapter (OnePlus)
- USB-C to USB-C cable
- The Pupil Invisible glasses, with a magnetically detachable front-view camera



Make sure the phone is charged. When charged, turn it on by holding the power button on the right side.

NOTE: Do NOT update the Android version on the phone itself (it should be 11). The phone is delivered with a compatible Android version, tested by the manufacturers. Should you accidentally upgrade to an incompatible Android version, you can rollback to a compatible version: https://docs.pupil-labs.com/invisible/data-collection/troubleshooting/#i-accidentally-updated-my-companion-device-to-an-incompatible-android-version

Connect the phone and the glasses using the cable. The USB-C port in the glasses is behind the right ear.

Performing measurement

Log in to the phone, using the passcode **** and connect it to Wi-Fi.

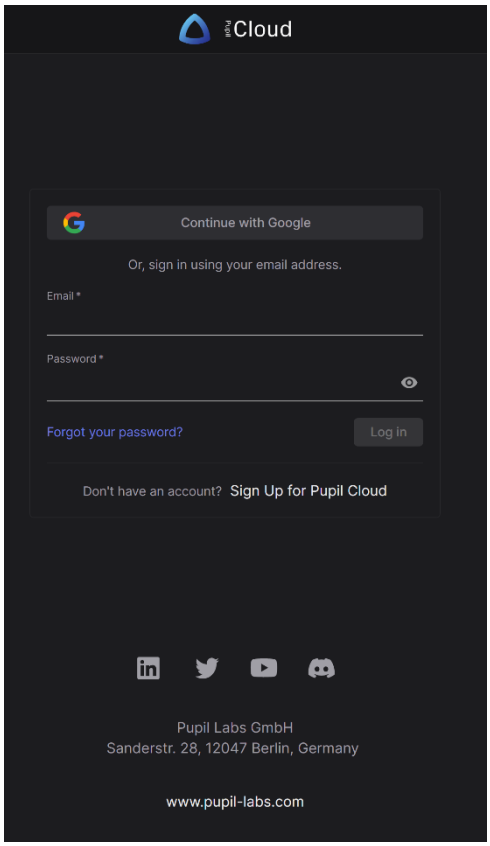
The following steps can also be seen in the screen recording demo.

- Open the settings, go to system → date & time, and toggle “Use network-provided time” off and back on.
- Open the Invisible Companion app.
- Bottom there is a preview button, this should give the feed from the front-view camera on your screen.
- Choose the wearer, or create a new one via the bottom-left menu.
- Start the recording (big red button).
- End the recording (same big red button).
- Click on the bottom-left menu, and go to recordings. It should be the top one. You can click it to view the recording. It will be automatically uploaded to the cloud.

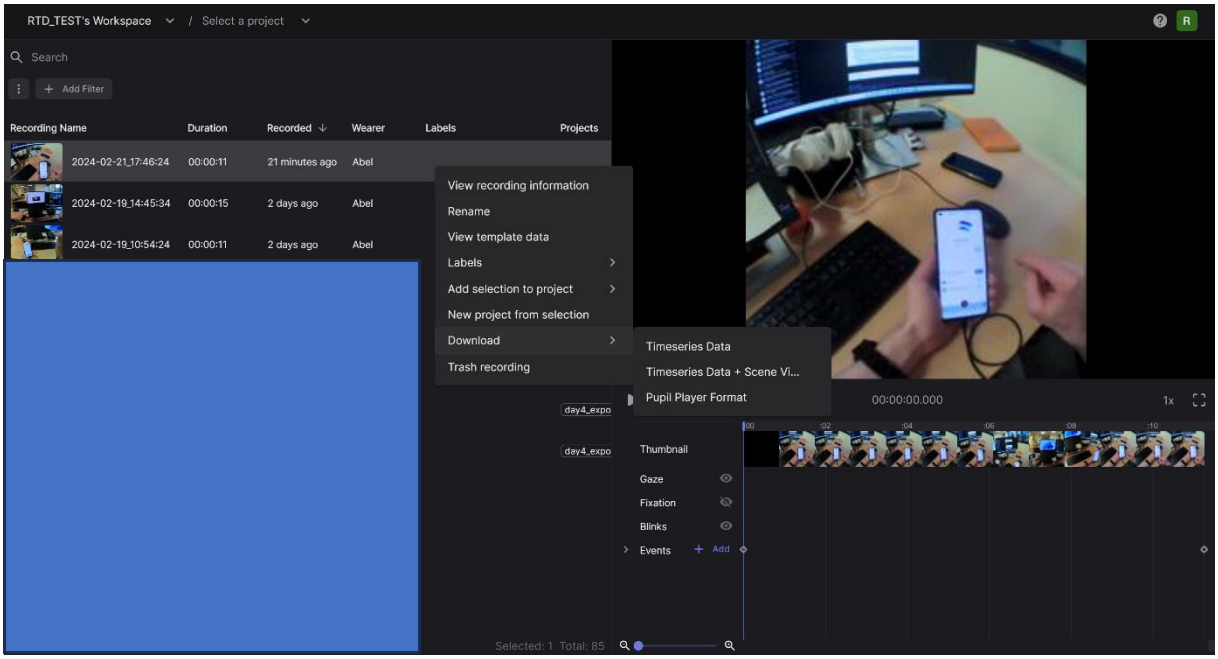
View and export recordings

On your computer, go to [https://cloud.pupil-labs.com/.](https://cloud.pupil-labs.com/)

Log in with the account that is coupled to the phone.



You will find the recordings. Right-click on the recording you're interested in and move the cursor to download to find the available data exports:



(The blue square is added for privacy reasons)

Downloading Timeseries Data + Scene Video will give you a .zip file with a folder that is titled the date and time of the recording. In this folder, you find the first-person video, and the CSV files of all the different data streams. The data streams are explained in more detail in the documentation of pupil-labs: https://docs.pupil-labs.com/invisible/data-collection/data-streams/

Downloads › Timeseries Data + Scene Video (7).zip › 2024-05-22_11-35-11-ad506fc1

Sort View Extract all ...

Name	Type	Compressed size	Password ...	Size	Ratio
3d_eye_states.csv	Microsoft Excel Comma S...	1 KB	No	1 KB	68%
blinks.csv	Microsoft Excel Comma S...	2 KB	No	9 KB	85%
events.csv	Microsoft Excel Comma S...	1 KB	No	1 KB	44%
f4bdc307_0.0-414.768.mp4	MP4 Video File (VLC)	189.380 KB	No	189.380 KB	0%
fixations.csv	Microsoft Excel Comma S...	55 KB	No	208 KB	74%
gaze.csv	Microsoft Excel Comma S...	2.756 KB	No	12.632 KB	79%
imu.csv	Microsoft Excel Comma S...	3.267 KB	No	15.439 KB	79%
info.json	JSON File	1 KB	No	1 KB	52%
saccades.csv	Microsoft Excel Comma S...	68 KB	No	231 KB	71%
scene_camera.json	JSON File	1 KB	No	1 KB	62%
world_timestamps.csv	Microsoft Excel Comma S...	56 KB	No	1.077 KB	95%

The CSV files already contain UNIX timestamps in nanoseconds! :)

F

Facial expression analyses

The PDF version of the analyses was included in overleaf. It starts on the next page. This appendix has been omitted in the public version.

G

Model Architecture

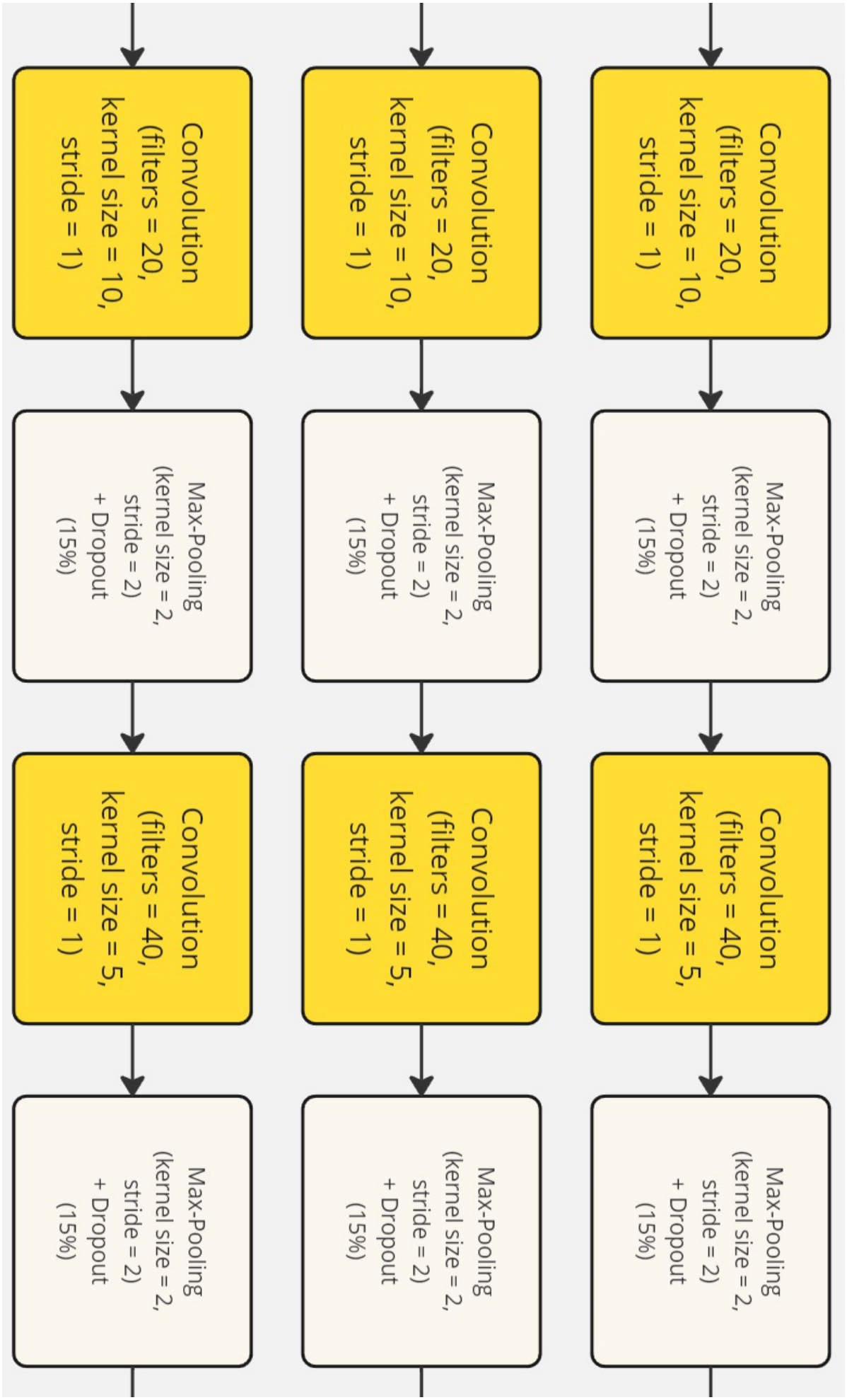


Figure G.1: Part 1/3 of the architecture.

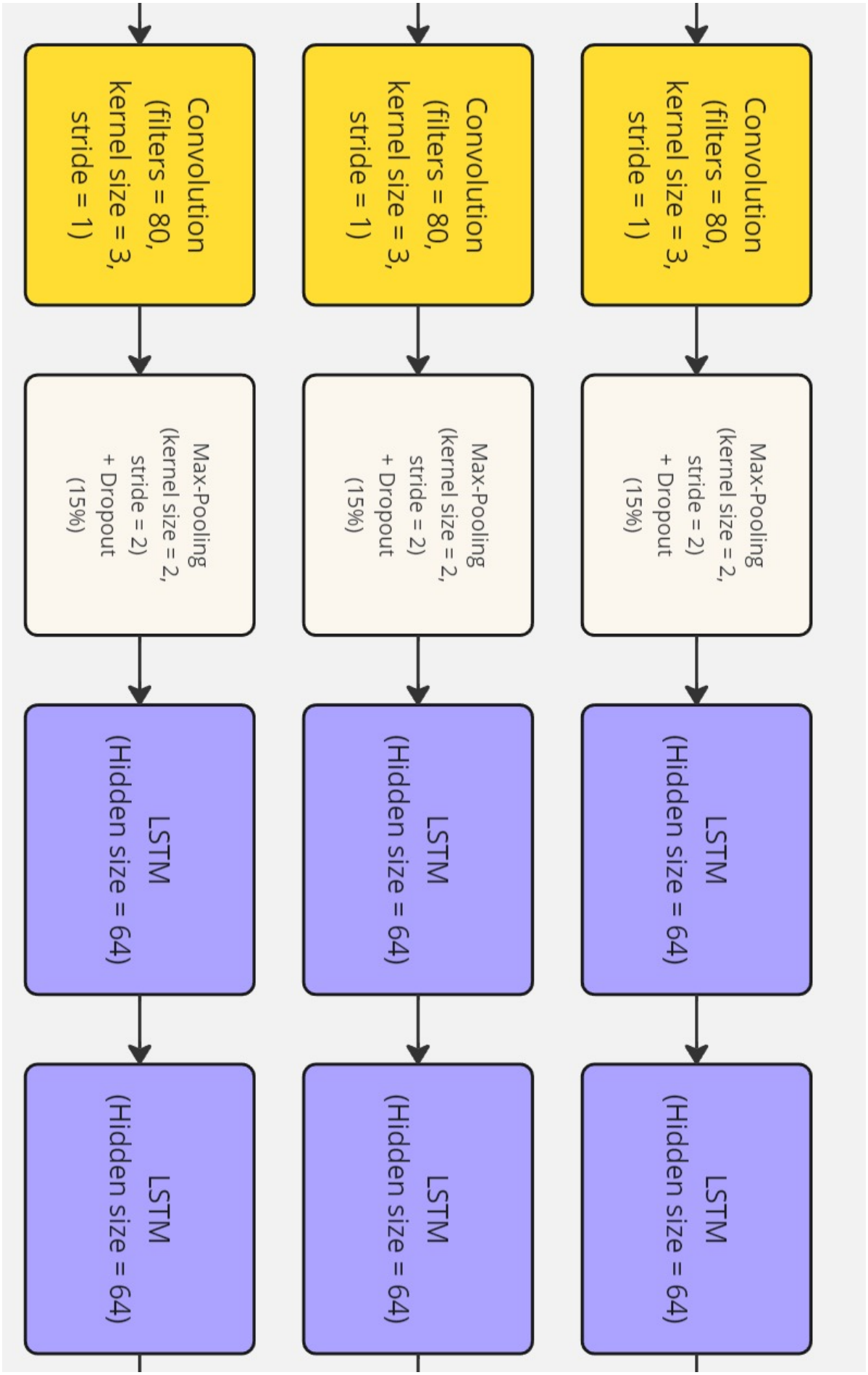


Figure G.2: Part 2/3 of the architecture.

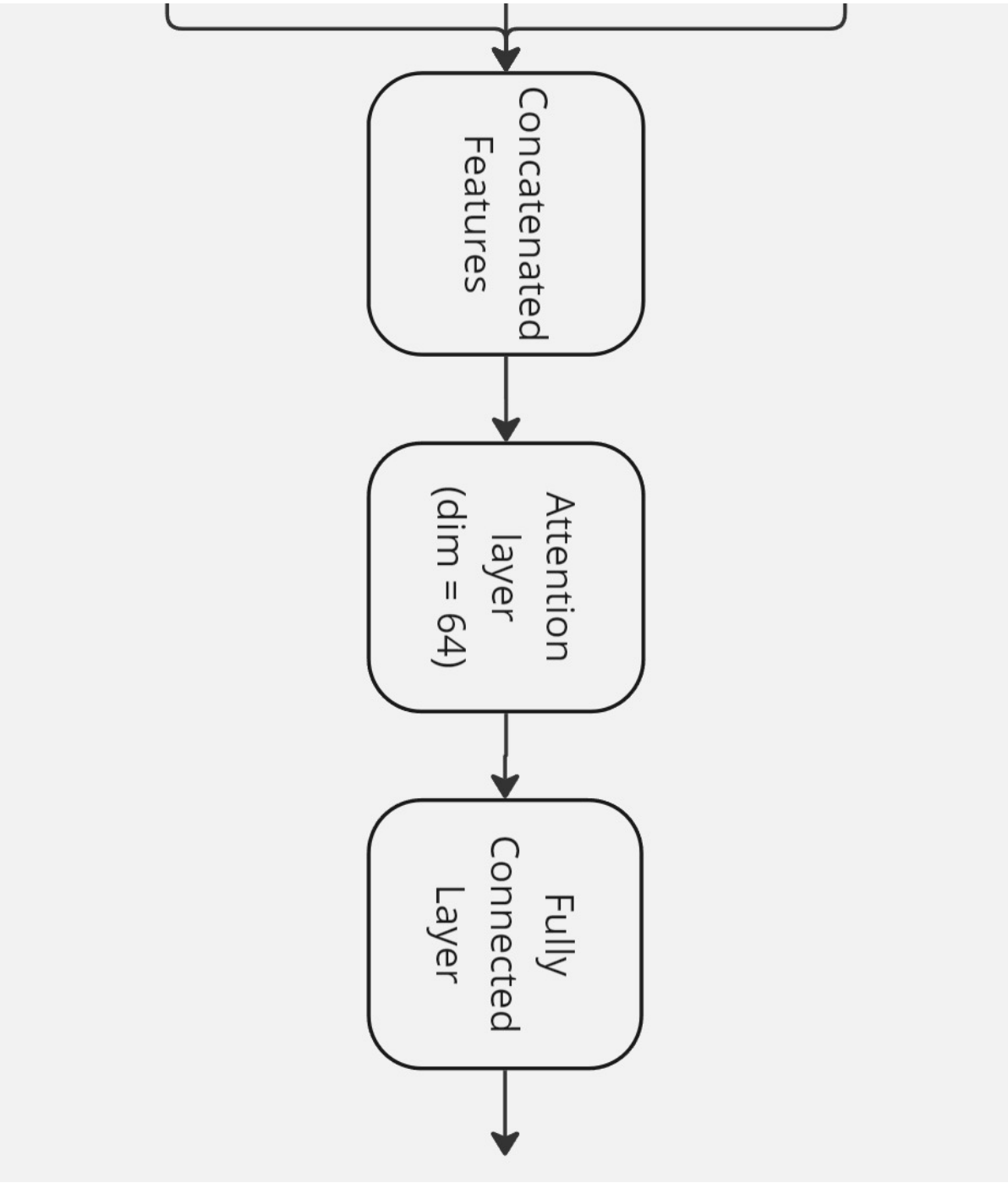


Figure G.3: Part 3/3 of the architecture.

H

Prediction plots

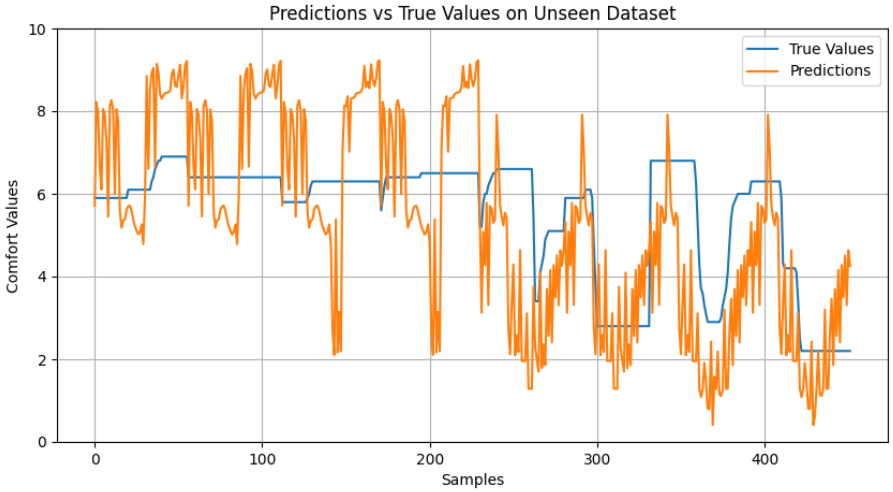


Figure H.1: Originally given subjective comfort ratings by participant 6 versus the prediction by the model. The model follows the trend with a correlation value of 0.62.

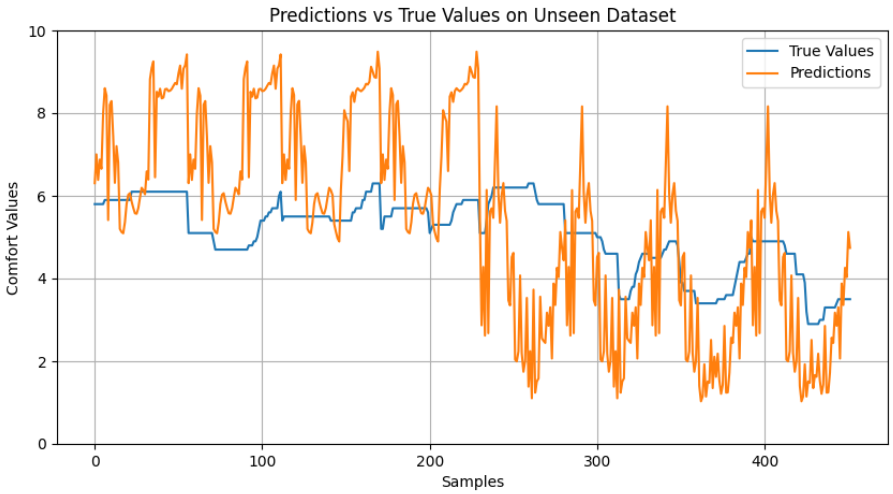


Figure H.2: Originally given subjective comfort ratings by participant 10 versus the prediction by the model. The model follows the trend with a correlation value of 0.62.

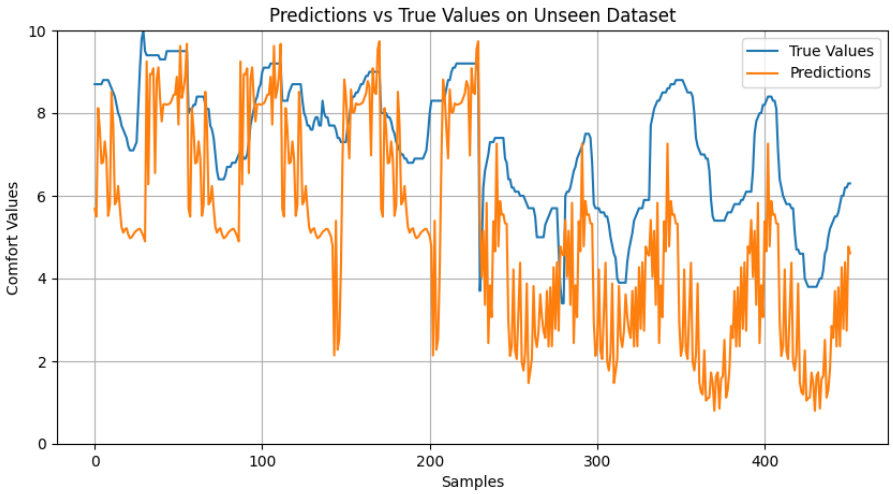


Figure H.3: Originally given subjective comfort ratings by participant 12 versus the prediction by the model. The model follows the trend with a correlation value of 0.74.

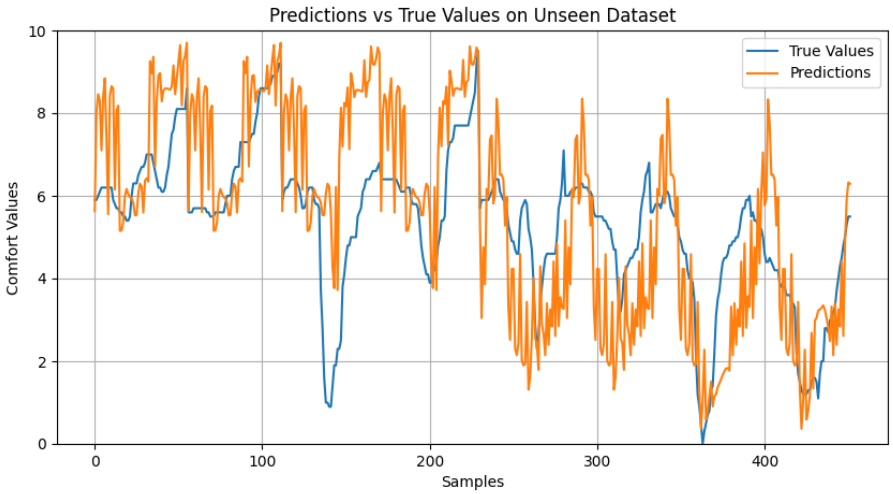


Figure H.4: Originally given subjective comfort ratings by participant 15 versus the prediction by the model. The model follows the trend with a correlation value of 0.68.

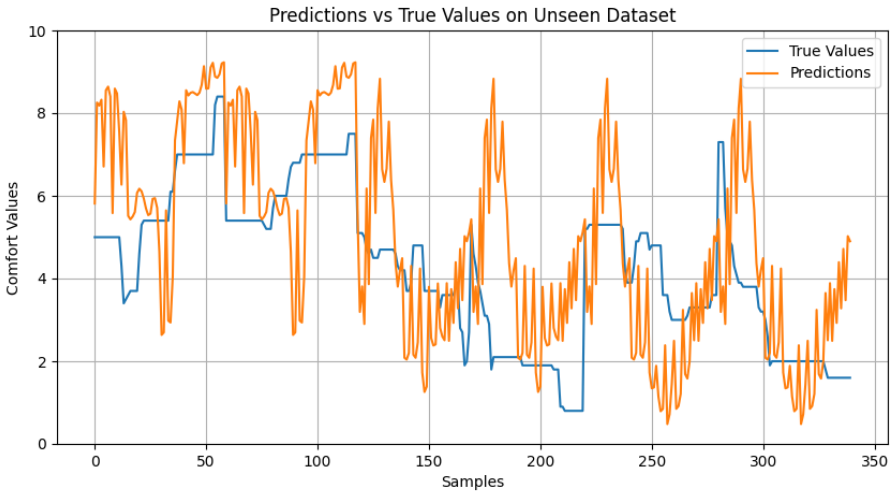


Figure H.5: Originally given subjective comfort ratings by participant 16 versus the prediction by the model. The model follows the trend with a correlation value of 0.60.

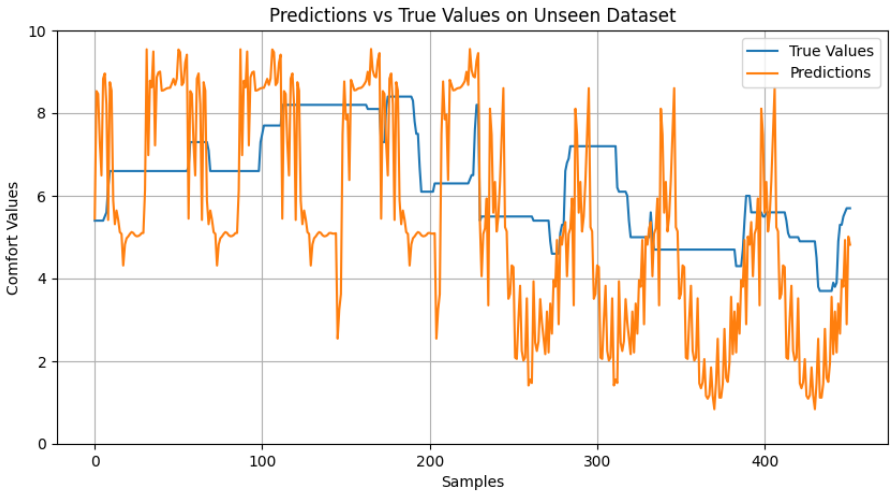


Figure H.6: Originally given subjective comfort ratings by participant 17 versus the prediction by the model. The model follows the trend with a correlation value of 0.55.

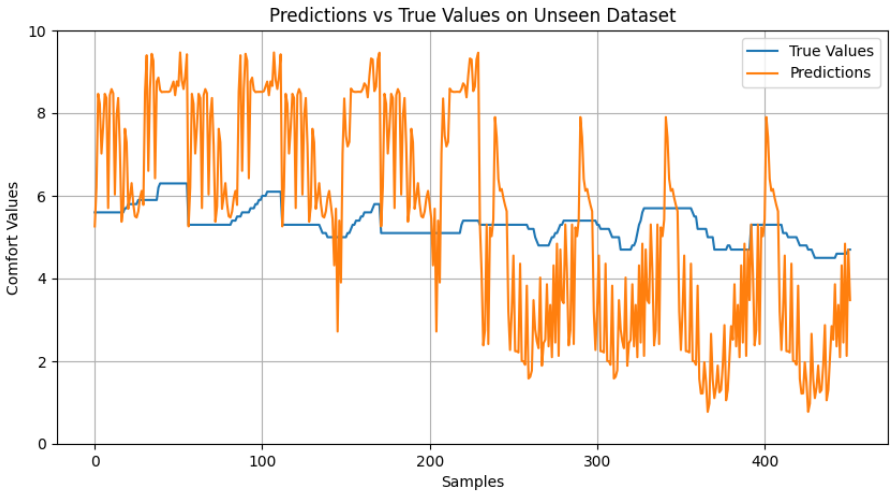


Figure H.7: Originally given subjective comfort ratings by participant 18 versus the prediction by the model. The model follows the trend with a correlation value of 0.59.

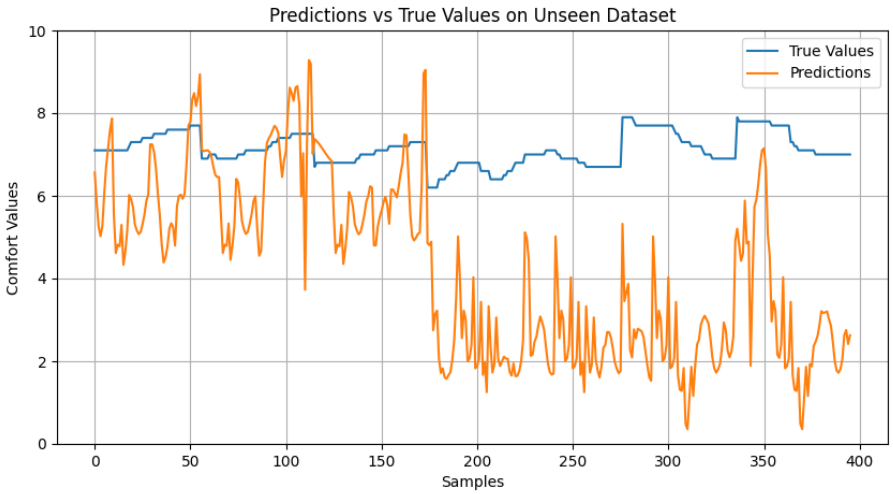


Figure H.8: Originally given subjective comfort ratings by participant 20 versus the prediction by the model. The model follows the trend with a correlation value of 0.28.

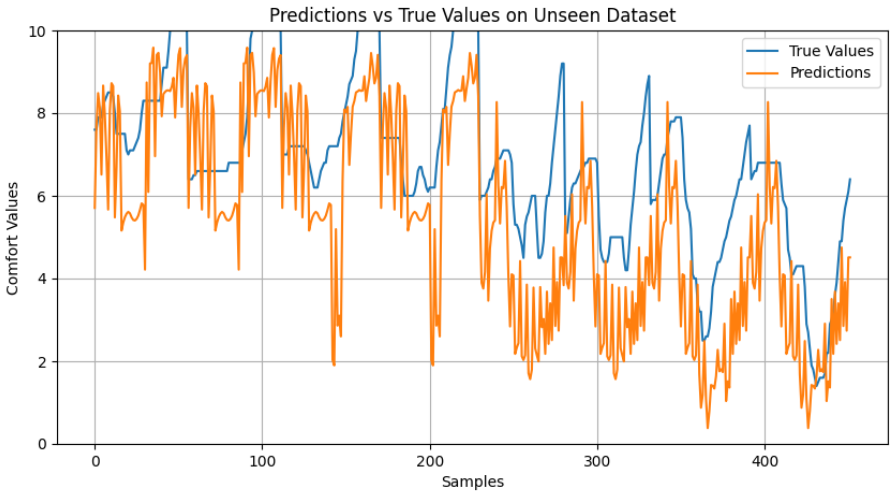


Figure H.9: Originally given subjective comfort ratings by participant 23 versus the prediction by the model. The model follows the trend with a correlation value of 0.80.

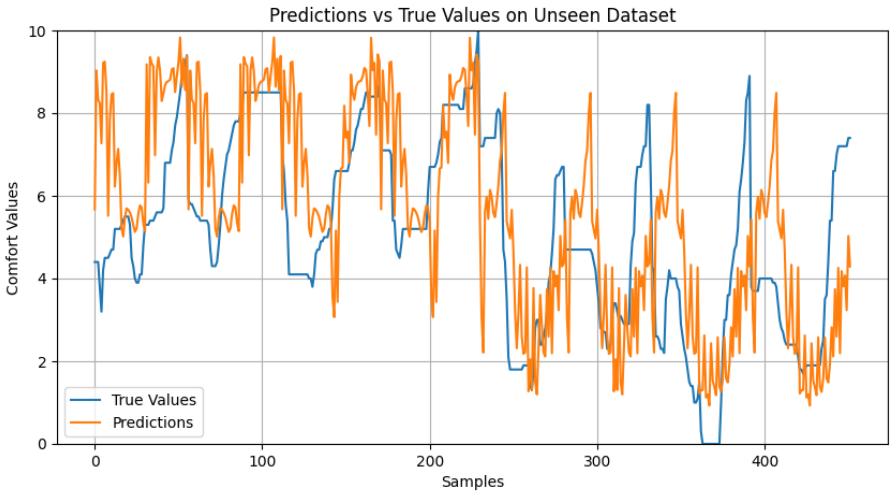


Figure H.10: Originally given subjective comfort ratings by participant 25 versus the prediction by the model. The model follows the trend with a correlation value of 0.64.

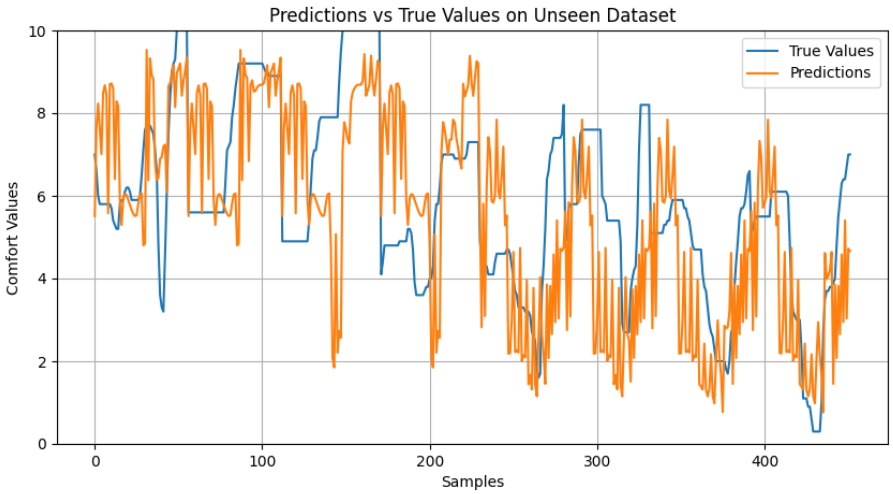


Figure H.11: Originally given subjective comfort ratings by participant 26 versus the prediction by the model. The model follows the trend with a correlation value of 0.58.

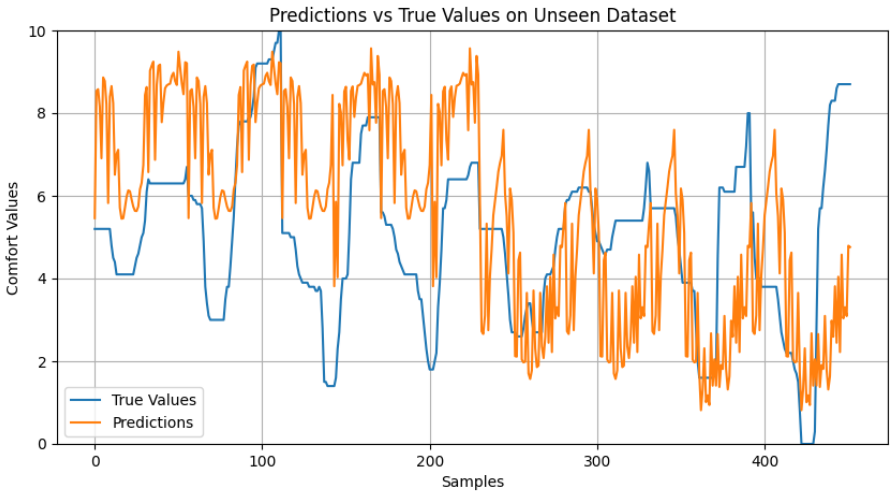


Figure H.12: Originally given subjective comfort ratings by participant 29 versus the prediction by the model. The model follows the trend with a correlation value of 0.41.

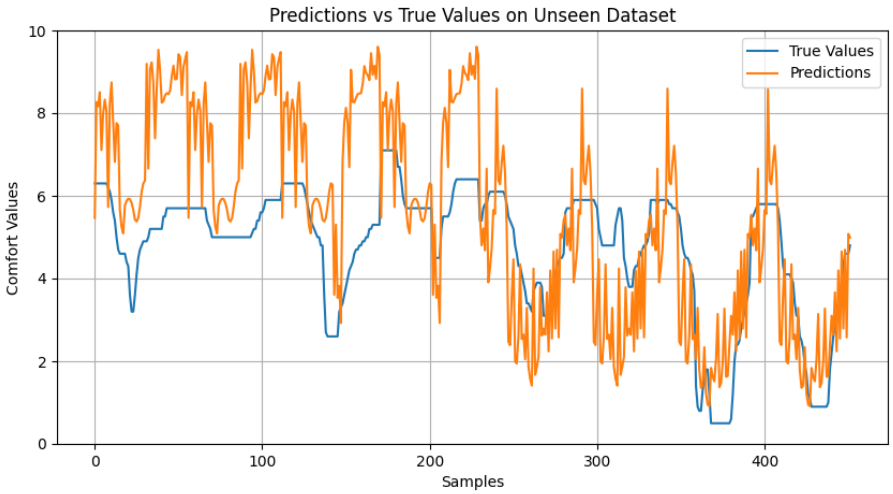


Figure H.13: Originally given subjective comfort ratings by participant 35 versus the prediction by the model. The model follows the trend with a correlation value of 0.66.

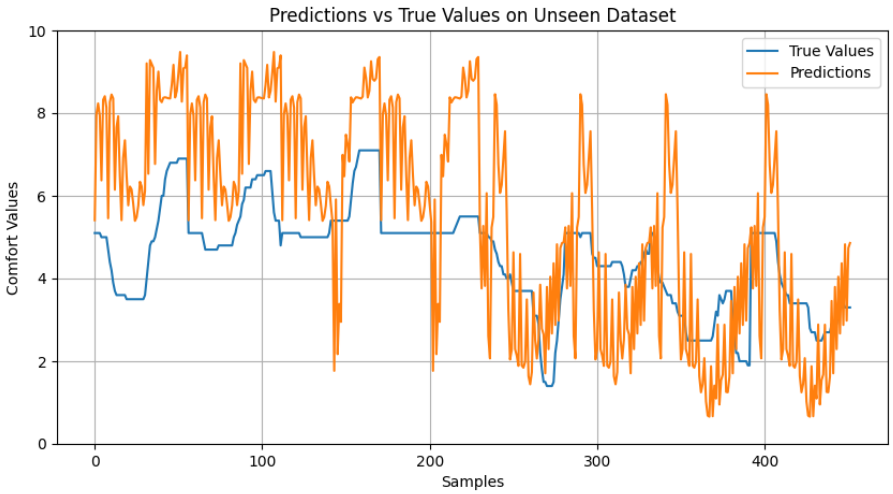


Figure H.14: Originally given subjective comfort ratings by participant 36 versus the prediction by the model. The model follows the trend with a correlation value of 0.71.